ARI Research Note 95-34

Development and Construct Validation of the Situational Judgment Test

Mary Ann Hanson and Walter C. Borman

Personnel Decisions Research Institutes, Inc.

for

Contracting Officer's Representative
Leonard A. White

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

Manpower and Personnel Research Division
Zita M. Simutis, Director

April 1995

United States Army
Research Institute for the Behavioral and Social Sciences

JGK

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

Research accomplished under contract
for the Department of the Army

Personnel Decisions Research Institutes, Inc.

Technical review by

Leonard A. White

NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1995, April	Final Apr 92 - Sep 92

4. TITLE AND SUBTITLE
Development and Construct Validation of the Situational Judgment Test (SJT)

5. FUNDING NUMBERS
MDA903-92-M-3490
62785A
791
2211
C07

6. AUTHOR(S)
Hanson, Mary Ann; and Borman, Walter C.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Personnel Decisions Research Institutes, Inc.
43 Main Street S.E.
Riverplace, Suite 405
Minneapolis, MN 55414

8. PERFORMING ORGANIZATION REPORT NUMBER
Institute Report #230

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARI Research Note 95-34

11. SUPPLEMENTARY NOTES
Prepared under project "Construct Validity of the Situational Judgment Test (SJT)"
Contracting Officer's Representative, Leonard White.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
--

13. ABSTRACT *(Maximum 200 words)*

This report describes the development of the Situational Judgment Test (SJT), the development and evaluation of basic SJT scores, explorations of the dimensionality of the SJT, and detailed investigations of the relationships between SJT scores and scores on temperament, cognitive ability, and other job performance measures. The SJT was developed to be a criterion measure of supervisory job knowledge and administered to over 1,000 second-tour Noncommissioned Officers (NCOs) in the U.S. Army. These data were used, along with several rational approaches, to explore the dimensionality of the SJT. Relationships between SJT total scores, several experimental SJT subscores, and scores on the other available measures were also examined, and structural modeling was used to test several hypotheses concerning reasons for some of the relationships that were found. Finally, conclusions were drawn, based on the results of these analyses, concerning what the SJT measures.

14. SUBJECT TERMS			15. NUMBER OF PAGES
Situational Judgment Test	Supervision	Project A	93
Job Knowledge Test	Leadership	Construct validation	16. PRICE CODE
Second-tour performance	Career force	Criterion measure	--

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

FOREWORD

The Situational Judgment Test (SJT) is a multiple choice, paper-and-pencil test that was developed as part of Project A to be a criterion measure of supervisory skill for first-line supervisors or Noncommissioned Officers (NCOs) in the U.S. Army. Initial data analyses were conducted for the SJT as part of Project A and the follow-on project, Building and Retaining the Career Force. Results showed that SJT scores are reliable and appropriately, yet not strongly, related to scores on the other measures of supervisory performance. Based on these results, it was concluded that the SJT may provide a significant amount of genuinely unique variance relative to leadership performance. However, additional research aimed at a better understanding of what the SJT measures was recommended.

The research described in this document provides additional information concerning the leadership constructs measured by the SJT and the potential utility of this test as a measure of supervisory performance. Analyses were conducted to explore the internal structure of the SJT and to provide information concerning the relationships between SJT scores, scores on other performance measures, amount of supervisory training and experience, and measures of ability and temperament. Results of these analyses provide insight concerning the leadership constructs measured by the SJT. In general, the SJT is probably best interpreted as a measure of supervisory job knowledge.

Based on these results, the SJT may be useful to the Army in several ways. First, results suggest that the SJT has potential for providing a valid indication of soldiers' leadership potential. In addition, a wide variety of situations and response alternatives were generated during the development of the SJT, and these materials have potential for use in developing instructional materials for supervisory training courses. Finally, the SJT may be useful in determining the effectiveness of supervisory training programs or in providing an assessment of supervisory training needs.

ACKNOWLEDGMENTS

The authors wish to thank Bob Cudeck at the University of Minnesota for his patience in dealing with our structural modeling questions and problems and his statistical wisdom that kept our analyses on track. We also extend heartfelt thanks to three Personnel Decisions Research Institutes staff members—Janis Houston, Leissa Nelson, and Cheryl Paullin—who carefully completed the arduous rating task that formed the basis for many of the analyses reported here. Without their conscientious efforts, much of this research would not have been possible.

DEVELOPMENT AND CONSTRUCT VALIDATION OF THE SITUATIONAL JUDGMENT TEST (SJT)

EXECUTIVE SUMMARY

Requirements:

The Situational Judgment Test (SJT) is a multiple choice, paper-and-pencil test that was developed as part of the Army's Project A to be a criterion measure of supervisory skill for first-line supervisors or Noncommissioned Officers (NCOs) in the U.S. Army. The purpose of this research was to bring together available information concerning exactly what is measured by this test and to conduct additional research concerning the internal structure of the SJT and its relationships with other measures to further our understanding of what the test measures.

Procedures:

The SJT was previously administered to a sample of more than 1,000 NCOs, along with a variety of other job performance measures and a temperament inventory. Armed Forces Qualification Test (AFQT) scores were also available for these soldiers. These data were used to further explore the internal structure of the SJT and the relationships between scores on the SJT and other available measures. The structure of the SJT was explored both empirically and rationally. Rational approaches involved collecting ratings of a variety of characteristics of the SJT items and response alternatives to identify important similarities and differences among the SJT items and response alternatives. Some of these ratings were used to develop several sets of experimental SJT subscales. Other ratings were used to explore the reasons some SJT response alternatives are more effective than others. This included determining which sources of power (e.g., coercive power, legitimate power) are more effective and which characteristics of the situations (i.e., item stems) impact the relative effectiveness of the various sources of power.

Relationships between SJT total scores, the experimental SJT subscores, and scores on the available temperament, cognitive ability (i.e., AFQT), and job performance measures were also examined. Job performance measures included several supervisory simulation or "role play" exercises, performance ratings made by the examinees' supervisors, and scores on technical job knowledge tests. Finally, structural model analyses were conducted to test a series of hypotheses

concerning the underlying reasons for the correlations between SJT scores and scores on other measures.

Findings:

All of the results of the present research are consistent with the interpretation of the SJT as a measure of supervisory job knowledge. For example, SJT scores are moderately correlated with scores on other measures of supervisory performance (e.g., the supervisory simulations). Also, soldiers with more supervisory training obtained significantly higher SJT scores. Correlations between the experimental SJT subscores and scores on other measures indicated that these subscales measure somewhat different but correlated aspects of supervisory job knowledge.

Research on the sources of power used in SJT response alternatives showed that, on average, the use of information power is the most effective and the use of coercive power is the least effective. In addition, certain characteristics of the item stems or situations—specifically the objective and direction (upward versus downward) of the influence attempt—affect the frequency with which the various sources of power occur in the SJT response alternatives and the effectiveness of each source of power when it occurs.

Results of the structural model analyses suggest that the SJT mediates the relationship between cognitive ability (i.e., AFQT scores) scores on the other supervisory performance measures. This provides further support for the notion that the SJT measures supervisory job knowledge; soldiers have to know what to do before they can do it effectively, and general mental ability would be expected to have an effect on supervisory performance through this learning process. In addition, supervisory experience and training appear to mediate the relationships of Dominance and Work Orientation with SJT scores, suggesting that more dominant, hard working soldiers are likely to obtain more supervisory experience and training which, in turn, leads to higher SJT scores.

Utilization of Findings:

The SJT is currently being used as a criterion measure in the Career Forces validation research. The subscales that were developed in this research will be useful in understanding how the SJT fits into models of second-tour soldier performance and may be useful in the development of criterion composites for validation purposes. In addition, because the SJT has been shown to measure supervisory job knowledge, it may be useful in assessing supervisory performance or potential for other purposes. For example, the SJT might be used to assess the

effectiveness of supervisory training programs or to make promotion decisions. Finally, results of this research provide detailed information concerning what is measured by certain subsets of SJT items and the reasons that some response alternatives are more effective than others. This information could be very useful in the development of other similar instruments in the future.

DEVELOPMENT AND CONSTRUCT VALIDATION OF THE SITUATIONAL JUDGMENT TEST (SJT)

CONTENTS

CONTENTS (Continued)

LIST OF TABLES

CONTENTS (Continued)

LIST OF FIGURES

DEVELOPMENT AND CONSTRUCT VALIDATION OF THE SITUATIONAL JUDGMENT TEST (SJT)

INTRODUCTION

Purpose of the SJT and a Brief Overview of Previous Research

The Situational Judgment Test (SJT) was developed, as part of a large selection and classification project for the U.S. Army called Project A (Campbell & Zook, 1991, provide an overview), to be a criterion measure of second-tour soldier performance. The main objectives in Project A were to validate the Armed Services Vocational Aptitude Battery (ASVAB) by collecting job performance data from rep-resentative samples of soldiers and to develop and evaluate new predictors of job performance.

Project A included an in-depth job analysis for nine representative second-tour soldier jobs, and the results showed that these jobs have performance requirements in both technical and supervisory areas (see Campbell, 1989 for details). This job analysis also showed that supervision and leadership represent a sizable portion of these soldiers' jobs. Analysis of the supervisory tasks these soldiers perform revealed nine supervision/leadership task clusters, and these task clusters are defined in Table 1. The methods for measuring supervisory performance in Project A were selected based on feasibility, cost, estimated construct validity, and appropriateness for job content. Three supervisory simulations (i.e., role play exercises) were developed to measure skills in two of the supervisory task clusters: counseling and training subordinates. Due to the costly and time-consuming nature of role play exercises, it was not feasible to develop this type of "hands-on" supervisory measure to tap additional supervision/leadership task clusters. Behaviorally anchored rating scales were also developed based on a critical incident job analysis, and three of these scales tapped supervisory aspects of the job. These behaviorally anchored scales were supplemented with seven additional rating scales based on the supervisory dimensions that were identified through the task analysis.

The goal in developing the Situational Judgment Test (SJT) was to construct a multiple choice, paper-and-pencil measure of supervisory skill. Because paper-and-pencil measures are more economical to administer than hands-on type measures, it was expected that this test could be used to measure a broader range of supervisory tasks than the supervisory simulations. The SJT was designed to measure the knowledge component of supervisory skill: knowing how to respond effectively in supervisory situations. Thus, it would be seen as a job knowledge test covering the supervisory part of the job. The development of the SJT has already been completed (Hanson & Borman, 1989; Campbell, 1991). For each of the 35 items on the SJT, soldiers read a description of a difficult supervisory situation, examine three to five possible responses to the situation, then select the most and the least effective response alternatives. The following example is representative of the kind of items that make up the SJT (this is not an actual SJT item):

Table 1

Supervision/Leadership Task Clusters from Second-Tour Job Analysis

1. Planning Operations. Activities that are performed in advance of major operations of a tactical or technical nature. That is, planning for, getting ready for, and developing orders for various kinds of team operations, whether it be combat, support, or technical operations. It is the activity that comes before actual execution out in the field or work place.

2. Directing/Leading Teams. The tasks in this category are concentrated in the combat and military police MOS. They involve the actual direction and execution of combat and security team activities. They occur out in the field and are heavily dependent on MOS-specific skills. Leading reconnaissance teams, setting up offensive and defensive positions, carrying out a fire mission, directing the clearing of mine fields, etc. would all be part of this category. They require "real time" decision making under pressure.

3. Monitoring/Inspecting. This cluster includes interactions with subordinates that seem to involve keeping an operation going once it has been initiated, such as checking to make sure that everyone is carrying out their duties properly, assisting people to overcome problems, making sure everyone has the right equipment, monitoring or evaluating the status of equipment readiness, supply levels, completeness of written reports, adequacy of current operating procedures, etc. This is a non-combat or non-crisis set of activities.

4. Individual Leadership. The content of the tasks in this cluster reflects attempts to influence the motivation and goal direction of subordinates by means of goal setting, interpersonal communication, sharing hardships, building trust, etc.

5. Acting as a Model. This dimension is not tied to a specific task content but refers to the NCO modeling the correct performance behavior whether it be technical task performance under adverse conditions, or exhibiting appropriate military bearing. The NCO sets the example.

6. Counseling. A one-on-one interaction with a subordinate during which the NCO provides support, guidance, assistance, and feedback on specific performance or personal problems that the soldier might be experiencing. It includes counseling on problems of a disciplinary nature.

7. Communication with Subordinates, Peers, and Supervisors. The tasks in this category deal with composing specific types of orders, briefing subordinates on things that are happening, and communicating information up the line to superiors, as with peers. Information is disseminated in both written and oral formats.

8. Training Subordinates. This very distinct cluster of tasks describes the day-to-day role of the NCO as a trainer for individual subordinates. When such tasks are being executed, they are clearly identified as instructional (as distinct from evaluations or disciplinary actions). They involve scheduling, planning, and conducting training.

9. Personnel Administration. This category is made up of "paperwork" or administrative tasks that involve actually doing performance appraisals, making or recommending various personnel actions, keeping and maintaining adequate records, and following standard operating procedures for Army personnel practices.

You are a squad leader on a field exercise, and your squad is ready to bed down for the night. The tent has not been put up yet, and nobody in the squad wants to put up the tent. They all know that it would be the best place to sleep since it may rain, but they are tired and just want to go to bed. What should you do?

a. Tell them that the first four men to volunteer to put up the tent will get light duty tomorrow.
b. Make the squad sleep without tents.
c. Tell them that they will all work together and put up the tent.
d. Explain that you are sympathetic with their fatigue, but the tent must be put up before they bed down.

The SJT has been administered to over 1000 second-tour soldiers. These data were used to develop scoring procedures for the SJT and to assess the psychometric characteristics of these scores (Hanson & Borman, 1990). Analyses have also been conducted that provide information concerning how the SJT fits into a model of second-tour soldier performance (Campbell & Zook, 1990).

Purpose of the Present Research

The purpose of the present research was to further assess the construct validity of the SJT, in other words to clarify and define exactly what is being measured by this test. The SJT was designed to measure supervisory job knowledge, but the results of previous research leave some unanswered questions concerning exactly what the SJT measures. Previous research does provide preliminary information concerning the construct validity of the SJT, and the most relevant portions are summarized in the present report. The present research involved more in-depth explorations of the content and dimensionality of the SJT. Relationships between SJT scores and scores on several temperament, cognitive ability, and job performance measures were also examined, and analyses were conducted aimed at better understanding the reasons for these relationships. Results of the present research are discussed in the context of previous research on the SJT and relevant portions of the literature. This report begins with a brief review of research on measures similar to the SJT.

Brief Review of Relevant Literature

Situational Judgment Tests

Situational judgment tests typically involve presenting respondents with realistic job situations, usually described in writing, and asking them to respond in a multiple-choice format regarding what should be done in each situation. Situational judgment tests have typically been developed by other researchers to predict job performance, especially for management and supervisory positions (e.g., Motowidlo, Bunnette & Carter, 1990; Mowry, 1964; Rosen, 1961; Tenopyr, 1969). An example of a situational judgment test designed to predict supervisory performance is the Leadership Evaluation and Development Scale (LEADS; Mowry, 1964), which is described by the authors as a measure of supervisory judgment. Tenopyr (1969) studied the concurrent validity of LEADS

in a sample of 126 production managers. These managers' LEADS scores correlated .36 with salary corrected for age and length of service and .25 with performance ratings by labor relations staff. LEADS also had a moderately high correlation with a test of verbal comprehension (.49), but the verbal comprehension test had lower correlations with corrected salary (.29) and the rating criterion (.08).

Mandell (1950) developed a similar test, the Administrative Judgment Test, and examined relationships between scores on this test, ratings by peers and supervisors, and pay grade in four relatively small samples (sample sizes ranging from 20 to 63). The median correlation between scores on the Administrative Judgment Test and these criteria was .51, and the Administrative Judgment Test also had moderately high correlations with scores on several mental ability tests (in the 50s and 60s). However, the mental ability tests that were included in this research had lower correlations with the criterion measure; their median validity was only .30. Similar concurrent validities have been obtained for several other supervisory or managerial situational judgment tests (e.g., Bruce & Learner, 1958; Rosen, 1961; Motowidlo, Dunnette, & Carter, 1990). Research investigating the longitudinal validity of situational judgment tests as predictors of supervisory or managerial job performance is not currently available in the literature.

Tacit Knowledge

Measures of "tacit knowledge" (e.g., Wagner & Sternberg, 1985) appear to be very similar to situational judgment tests. These tests present respondents with broad descriptions of situations and ask them to rate the importance of each in a list of possible behaviors for reaching the described goals. Measures of tacit knowledge do not directly ask respondents to choose one of the possible "responses", as do situational judgment tests, but by rating the importance of each behavior respondents are providing similar information. Research has shown that tacit knowledge correlates significantly with some measures of occupational success including salary increases, performance ratings, and expert versus novice status (Wagner & Sternberg, 1985).

Written Simulations

Written simulations have been used as measures of professional knowledge in several different fields, including law and medicine. These tests differ from situational judgment tests in that they typically employ a branching format; each response to a realistic job situation leads to more information about the situation, and the respondent is again asked to choose among a new group of response alternatives. Because written simulations are used as criterion measures, available research on the usefulness of these tests provides information concerning the potential usefulness of situational judgment tests as a criterion measures.

Much of the available research on written simulations supports their construct validity as measures of professional knowledge. However, this research varies widely in quality, and the results to date are far from conclusive. Comparisons of written simulation scores obtained by different groups of respondents have generally shown that when the groups are fairly distinct in

terms of training and experience (e.g., students versus professionals), differences are significant and in the expected direction. This finding is consistent across several content areas including legal simulations (e.g., Alderman, Evans, & Wilder, 1981) and medical simulations (e.g., McGuire & Babbott, 1976). Research on relationships between written simulation scores and other measures of job performance has obtained mixed results. While some researchers have found significant relationships between written simulation scores and other measures such as supervisory ratings and scores on high fidelity simulations, other researchers have conducted similar research and failed to find the expected relationships (Smith, 1983; Brull, 1981).

Issues in Measures of Aptitude versus Achievement

Situational judgment tests have been used primarily as predictors of job performance in the past, so the development of a situational judgment test to be used as a criterion measure warrants some explanation. The distinction between situational judgment tests used as predictor measures and their use as job performance measures may be viewed as similar to the distinction between aptitude tests and achievement tests. Angoff and Johnson (1988) summarize several generally accepted distinctions between the concept of aptitude and that of achievement. First, aptitude tests draw their items from a wide range of human experience, but material for achievement tests is necessarily more circumscribed. Second, aptitudes are expected to develop and change slowly, primarily as a consequence of general life experiences, while achievement typically increases rapidly as a result of exposure to information from the relevant content area. Finally, aptitudes are expected to resist short term efforts to hasten their growth, and achievement is expected to be susceptible to such efforts. It is likely that, as Humphreys (1974) proposed, tests of aptitudes and achievement actually fall on a continuum, and these definitions represent the extremes of this continuum.

Based on these distinctions, the supervisory knowledge or skill assessed by most situational judgment tests might be seen as having aspects both of an aptitude and of achievement. Some people may be able to answer questions concerning how to handle supervisory situations on the basis of common sense or general life experiences, while others obtain this ability through training. In addition, some situational judgment test items are likely to tap achievement more than others; for example, items concerning organizationally mandated methods for handling specific situations (e.g., refer alcohol abusers to a counselor) would clearly be achievement related. The test used in the present research -- the SJT -- was designed to be a criterion measure of supervisory job knowledge, so it is intended to measure achievement. To the extent possible, SJT items were written to tap the achievement component of supervisory knowledge or skill. However, some people are likely to acquire knowledge concerning how to effectively handle supervisory situations through general life experiences, so it is unlikely that the SJT is a pure measure of achievement. Due to the nature of supervisory tasks, it is likely that any job knowledge test in the supervisory domain will necessarily have some characteristics of an aptitude test.

The degree to which the knowledges and skills assessed in a situational judgment test are learned on the job is central in determining whether it is

appropriate as a predictor or as a criterion measure. If these knowledges and skills can be easily picked up on the job or learned in training that an applicant is likely to receive, situational judgment tests may be less useful as predictors. If on the other hand situational judgment test is to a large extent measuring an aptitude or ability, it is probably not appropriate as a criterion measure. Because it is likely that what is measured is in fact somewhat like an ability and somewhat like achievement, situational judgment tests may under different circumstances be useful for both purposes.

DEVELOPMENT OF THE SJT

The target population for the SJT is second-tour soldiers, and these soldiers are predominantly corporals, specialists 4/5, or sergeants working in beginning supervisory positions. Development of the SJT involved asking groups of Non-Commissioned Officers (NCOs) similar to the target sample to describe, in writing, a large number of difficult but realistic situations that Army first-line supervisors face on their jobs. Once a large number of these situations had been collected, a wide variety of possible actions (i.e., response alternatives) for each situation were gathered, and ratings of the effectiveness of each of these actions were collected from both experts (senior NCOs) and members of the target group of first-line supervisors. These effectiveness ratings were used to select situations and response alternatives to be included in the SJT. The effectiveness ratings from the senior NCOs (i.e., experts) were also the basis for the development of SJT scoring procedures. Details of the test development work appear below.

SJT Item and Response Alternative Development

Participants in the workshops to develop situations were 40 NCOs. Some were members of the target population (first-line supervisors) and some were NCOs who directly supervised soldiers in the target population. Eight workshops were conducted, each at a different Army post. A variation of the critical incident technique (Flanagan, 1954) was used to collect situations to be used as item stems. Workshop participants were asked to describe difficult supervisory situations that they or their peers had experienced as first-line supervisors in the Army. They were given training concerning how to write "good" situations based on the following criteria for a good situation:

1. It is challenging (i.e., difficult).
2. It is realistic.
3. It provides sufficient detail to help the supervisor make a choice between possible actions.
4. There is a correct way to respond, or at least some responses are better than others.
5. A response to the situation can be communicated in just a few sentences.

Participants were encouraged to write situations that tapped all nine categories of supervisory performance that had been identified in the task analysis described previously. Even so, there were some categories for which few if any situations were written (e.g., Acting as a Model, Training Subordinates), probably because they do not lend themselves well to this type of test format. A total of over 400 situations were generated in these workshops and then edited by research staff. In later workshops (described below), the NCOs who generated response alternatives reviewed these situations to ensure that they met the criteria listed above. Situations that could not be rewritten to meet the criteria were dropped. The end result was a total of 300 situations that met the criteria.

Fifty-two NCOs participated in the workshops to develop response alternatives, including both NCOs from the target population and supervisors of these

NCOs. Seven workshops were conducted, each at a different Army post. The goal in developing these response alternatives was to obtain a comprehensive list of responses likely to be chosen by first-line supervisors in the Army that also represent a variety of different levels of effectiveness. To accomplish this, workshop participants were presented with the situations and asked to write, in two or three sentences, what they would do to respond effectively in each situation. Most of these NCOs also participated in small group discussions, and this exercise often generated additional response alternatives. In addition, the NCO who originally wrote each situation also described what he or she would do in that situation. A total of approximately 15 responses were collected for each situation. All of these responses were content analyzed by research staff and collapsed when redundancies were noted, resulting in from four to ten response alternatives per situation with an average of about six.

One issue that surfaced during the development of the SJT concerned the wording of the questions or item stems. Because the SJT is intended to be a measure of job performance, the behavior of interest is what respondents actually would do in each situation. However, if asked "what would you do?" some respondents might indicate what they would actually do even if they recognize that a different response option would be more effective, while others might try to choose the response option that is most effective regardless of what they would actually do. In an effort to avoid this problem and to standardize the expected response set, the SJT instructions were written to ask respondents what should be done to respond effectively to each situation.

A second issue was the substantial amount of reading involved in responding to SJT items. There was some concern that the low average reading ability of the target NCO job incumbents would result in the SJT functioning merely as a measure of reading ability, so a reading level analysis was conducted using the FOG index (Gunning, 1952). Results indicated that the reading level required to complete the SJT is about the seventh grade. Because this is a very low reading level and the SJT is not a speeded test, it is unlikely that reading ability has a very strong effect on SJT scores.

Selection of the Final Set of SJT Items

From the sample of 300 "good" situations, 180 of the most promising were selected based on Army subject matter expert and researcher judgments concerning how well the situations met the criteria for a good situation, whether there were adequate numbers of plausible response alternatives, and how well the final sample of situations covered the supervisory knowledge domain. For each of these 180 situations, information concerning the effectiveness of the various response alternatives was then collected from two groups; a group of the target population NCO job incumbents and a group of the most effective senior NCOs in the Army. The target NCO sample was 344 second-tour soldiers who were participating in a field test of a variety of job performance measures at several different Army posts in the United States and Europe. The "expert" sample was a group of about 90 senior NCOs who were students and instructors at the United States Army Sergeants Major Academy (USASMA). These latter NCOs were some of the highest ranking enlisted soldiers in the Army. They all had extensive experience as Army supervisors, with an average of over 15 years in supervisory positions, so they were in an ideal position to

provide information about the actual effectiveness of the various SJT response alternatives.

For each SJT situation, both groups of respondents were asked to make two kinds of judgments. First, they were asked to indicate which response alternative was most effective and which was least effective. Then they were asked to rate the effectiveness of each response alternative on a seven point scale (where one was the least effective and seven was the most effective). Generally the alternatives that each respondent considered the most and least effective could be deduced from their ratings, but when several response alternatives were given the same effectiveness ratings, the judgments of most and least effective were useful. Because there were still 180 situations (i.e., items) being considered at the time these data were collected, each NCO rated the response alternatives for only a subset of the items. This resulted in about 25 expert NCO and about 45 incumbent NCO responses per situation.

Items (situations) for the SJT were selected based on these data. First, a subset of the situations were identified that met two criteria: (1) the expert group had good agreement concerning the most effective response in that situation, and (2) the situation was difficult for the incumbents (i.e., agreement was substantially lower than in the expert group). Sixty nine of the 180 situations (about 40%) met these two criteria. From this group, 35 items were chosen for the final test based on the extent to which they were judged to tap an important aspect of supervisory knowledge. From three to five of the best response alternatives were retained for each of these items.

After completing their rating task, the expert NCOs who participated in the development of the SJT were also asked to answer several questions concerning their opinions about the SJT. Eighty eight percent thought that the SJT was a fair way to assess knowledge of supervisory practices, and eighty nine percent said they would have more confidence in an NCO who obtained a relatively high score. All of them thought that the situations were at least somewhat realistic. Based on these results, it appears that the SJT has excellent face validity.

ADMINISTRATION OF THE SJT TO THE CVII SAMPLE AND BASIC ANALYSES

This section describes the administration of the Situational Judgment Test (SJT) to the Project A Concurrent Validation second-tour (CVII) sample, and analyses of these data to develop the basic SJT scores. These data were collected as part of a larger data collection effort that involved the administration of a variety of other job performance measures (see Campbell, 1991). Project A was the first phase of a two-phase program, and these data analyses were actually conducted as part of the second phase, Building and Retaining the Career Force (see Campbell & Zook, 1990). Because the SJT had not been thoroughly field tested prior to administration to the CVII sample, the CVII data collection was considered a field test of the SJT. There were two major objectives in the basic analyses of the SJT data from the CVII sample. The first objective was to examine and evaluate the psychometric properties of this instrument. The second objective was to develop one or more SJT scores to be used in the modeling of second-tour performance.

Administration Procedures

The SJT was administered to a total of 1049 soldiers in the CVII sample. Eleven percent of these soldiers were female, and the racial breakdown was as follows: 56 percent white, 33 percent black, and six percent Hispanic (the remainder reported "other"). These soldiers were sampled from ten different Army posts in the United States and several sites in Europe (USAREUR). A more detailed description of this sample is provided by Campbell and Zook (1990).

For each of the 35 SJT items, soldiers were asked to read the description of a supervisory situation, examine the possible responses, and select the most and least effective response alternatives. They were told to mark an "M" next to the response alternative they believed was the most effective and an "L" next to the response alternative they believed was the least effective. A detailed description of the administration procedures for the entire CVII data collection effort is available in Campbell and Zook (1990).

Data Screening and Scoring

Data Screening and Frequency Counts

These data were first screened for invalid and incomplete data. While the majority of the 1049 inventories were filled out completely and correctly, some data clean-up was required. Then, because a certain amount of valid data is necessary to compute reliable scores, SJT scores were computed only for soldiers who had valid item-level data for at least 90 percent of the items. A total of 1025 soldiers had valid "M" responses for at least 90 percent of the items, 1007 had valid "L" response for at least 90 percent of the items, and 1007 had both valid "M" and valid "L" responses for at least 90 percent of the SJT items.

Frequency counts of the number and percentage of respondents choosing each SJT response alternative were conducted to determine whether there was variability in the answers chosen by respondents in this sample. Because the SJT items are in a multiple-choice format, it is conceivable the correct answer is

obvious (i.e., the test is too easy). If this is the case, it would be impossible for SJT scores to discriminate among these soldiers. Results of these frequency counts showed that the SJT item-level responses from the CVII sample were distributed quite well across the response alternatives for each item. For example, the percentage of respondents indicating that the most frequently chosen response alternative for an item was the most effective ranged from 32 to 74, with a median of 46 percent. This suggests that the correct responses to SJT items were not at all obvious to the soldiers in this sample.

Development of Scoring Procedures

Several different approaches to scoring the SJT were explored. The most straightforward was a simple number correct score. For each item, the response alternative that was given the highest mean effectiveness rating by the USASMA experts was designated the "correct" answer. Respondents scores were simply the number of items for which they selected this "correct" response alternative as the most effective. The second scoring procedure involved weighting each response alternative soldiers selected as most effective by the mean effectiveness rating given to that response alternative by the expert group. This approach gives respondents more credit for choosing "wrong" answers that are relatively effective than for choosing wrong answers that are very ineffective. These item-level effectiveness scores were then averaged to obtain an overall effectiveness score for each soldier. Averaging these item-level scores instead of simply summing them places respondents' scores on the same one to seven effectiveness scale as the experts' ratings and does not penalized respondents for missing data.

Because the experts' ratings were used as item weights, the level of agreement among the experts concerning these effectiveness ratings was assessed to determine whether such weights would be adequately reliable. An intraclass correlation was computed for each of the 35 items included in the SJT across all of the experts who had rated that item. These analyses assumed that the experts were a random sample from a larger population (Shrout & Fleiss, 1979; Case 2). Reliabilities are reported here for the mean effectiveness ratings across all judges (between 20 and 26 judges for each item), which were used as the weights in scoring. These intraclass correlations ranged from .32 to .98 with a median of .93, and for 75 percent of the items the intraclass correlations were .82 or higher. This level of agreement was judged to be adequately high.

Scoring procedures based on respondents' choices for the least effective response to each situation were also explored. The ability to correctly identify the most ineffective response alternatives might be seen as an indication of a respondent's ability to avoid these very ineffective responses or in effect to avoid "screwing up". As with the choices for the most effective response, a simple number correct score was computed: the number of times each respondent correctly identified the response alternative that the experts rated the least effective. In order to differentiate this score from the number correct score based on choices for the most effective response, this score is referred to as the L-Correct score, and the score based on choices for the most effective response (described previously) is referred to as the M-Correct score. Another score was computed by weighting respondents' choices for the

least effective response alternative by the mean effectiveness rating for that response alternative, and then averaging these item-level scores to obtain an overall effectiveness score based on choices for the least effective response alternative (low scores are "good"). This score is referred to as L-Effectiveness, and the parallel score based on choices for the most effective responses (described previously) is referred to as M-Effectiveness.

Finally, a scoring procedure that involved combining the choices for the most and the least effective response alternative into one overall score was also explored. For each item, the mean effectiveness of the response alternative each soldier chose as the least effective was subtracted from the mean effectiveness of the response alternative they chose as the most effective. Because it is actually better to indicate that less effective response alternatives are the least effective, this score can be seen as a sum or composite of the two effectiveness scores described previously (i.e., subtracting a negative number from a positive number is the same as adding the absolute values of the two numbers). These item-level scores were then averaged together for each soldier to generate yet another score, and this score is referred to as M-L Effectiveness.

Descriptive Statistics for the Five Scoring Procedures

Descriptive statistics and an estimate of internal consistency reliability (coefficient alpha) were computed for the scores obtained from each of the five scoring procedures. The intercorrelations among these five scores were also computed. Finally, item analyses were conducted for each of the scoring procedures. Item-total correlations were computed for all five scoring procedures, and the proportion of the sample answering each item correctly was also computed for the M- and L-Correct scoring procedures.

Table 2 presents the mean score in this sample for each of the five scoring procedures. The maximum possible for the M-Correct scoring procedure is 35 (i.e., all 35 items answered correctly), but the maximum score obtained by soldiers in this sample was only 27, and the mean score was only 16.25. The mean number of least effective response alternatives correctly identified by this group was only 14.86. Clearly the SJT was difficult for this group of soldiers. Table 2 also shows the standard deviation, the minimum and maximum scores obtained, and the internal consistency reliability for each of the five scoring procedures. The internal consistency reliabilities for all five scoring procedures are quite high. The most reliable score is M-L Effectiveness, probably due to the fact that this score contains more information than the other scores (i.e., choices for both the most and least effective response alternative).

Table 3 presents the intercorrelations among scores obtained using the five different scoring procedures. These intercorrelations range from moderate to very high. Correlations between scores that are based on the same set of responses (e.g., M-Correct and M-Effectiveness) are higher than correlations between scores that are based on different sets of responses (e.g., M-Correct and L-Correct). The negative correlations between the L-Effectiveness score and the other scores are due to the fact that lower L-Effectiveness scores are actually better. The high (negative) correlation between M-Effectiveness and

Table 2

Situational Judgment Test (SJT) Means, Standard Deviations, and Internal Reliabilities

Scoring Procedure	N	Max. Score	Min. Score	Mean	SD	Coefficient Alpha
M-Correct (# of correct "Most" responses)	1025	27	3	16.52	4.29	.60
M-Effectiveness (mean eff. scale value for "Most" responses)	1025	5.65	3.66	4.91	.34	.68
L-Correct (# of correct "Least" responses)	1007	25	2	14.86	3.86	.57
L-Effectiveness (mean eff. scale value for "Least" responses)	1007	2.90[1]	4.84[1]	3.54[1]	.31	.68
M-L Effectiveness (mean of eff. scale value for "Most" responses minus eff. scale value for "Least" responses)	1007	2.57	-.77	1.36	.61	.75

[1] Low scores are "better;" mean effectiveness scale values for L-responses should be low.

Table 3

Situational Judgment Test (SJT) Score Intercorrelations for Five Basic Scoring Procedures

	M-Eff.	L-Correct	L-Eff.	M - L Eff.
M-Correct	.94	.52	-.64	.86
M-Eff.	---	.59	-.70	.93
L-Correct	---	---	-.86	.78
L-Eff.	---	---	---	-.92
M - L Eff.	---	---	---	---

Note. Sample sizes range from 1007 to 1025.

L-Effectiveness seems to indicate that these two scores are reflecting very similar or highly related constructs.

Table 4 shows the median and range of the 35 item-total correlations obtained for each of the five scoring procedures. These correlations are generally moderate, although there is a great deal of variability across items. As would be expected, the scoring procedures that yield more internally consistent scores also tend to have higher item-total correlations. For three of the SJT items the item-total correlations were extremely low for at least one scoring procedure (ranging from .00 to .09). Scores were recomputed for each of the five scoring procedures with these suspect items excluded, but the internal consistency reliabilities increased by only about .01 when these items were removed. Therefore, all of the SJT items were retained for the remainder of the analyses.

The proportion of the sample answering each item correctly could only be computed for the M- and L-Correct scoring procedures. There was a great deal of variability in this index of item difficulty across the 35 SJT items. Some items were answered correctly by less that 25 percent of the sample while others were answered correctly by up to 74 percent of the sample. This large range of item difficulties is likely to be useful in discriminating among respondents across the entire range of SJT scores.

Table 4

Summary of Situational Judgment Test (SJT) Item Analysis Results

	Item-Total Correlations		Proportion Answering Items Correctly	
Scoring Procedure	Range	Median	Range	Median
M-Correct	.03 to .47	.25	.24 to .74	.44
M-Effectiveness	.06 to .51	.29	---	---
L-Correct	.05 to .47	.21	.15 to .71	.40
L-Effectiveness	.00 to .54	.27	---	---
M-L Effectiveness	.05 to .52	.33	---	---

Based on the descriptive statistics presented here, the M-Correct and L-Correct scores appear to have less desirable psychometric characteristics than the scores obtained using the other three scoring procedures. Further, the M-L Effectiveness score is the most reliable and, based on its high correlations with both the M- and the L-Effectiveness scores, appears to provide an adequate summary of the information contained in the SJT responses. Thus, the remainder of the analyses presented in this report focus on the M-L Effectiveness score (which will be referred to as the SJT Total Score) .

Subgroup Differences in SJT Scores

Descriptive statistics for the SJT Total Score (i.e., M-L Effectiveness score) were computed separately for males and females and for several different racial subgroups. Descriptive statistics were also computed separately for soldiers from combat and non-combat MOS, and for soldiers from each of the nine MOS included in the present research. These latter analyses provide some information concerning whether the SJT is an equally appropriate measure of supervision for all nine MOS included in the present research. Some of the participants in the SJT development workshops reported that supervision in combat MOS is somewhat different from supervision in non-combat MOS. For example, some of them reported that supervisors in combat MOS are expected to take a stricter approach to subordinate misconduct. If the "correct" answer to SJT items varies by MOS, this may be reflected in differences in the mean scores of soldiers from different MOS.

Table 5 shows that females tend to score higher on the SJT than males, about a third of a standard deviation higher. Analysis of variance revealed that this difference is significant, but does not account for a great deal of the variance in SJT scores. In addition, blacks scored lower than whites and Hispanics, on average a little more than a third of a standard deviation lower. Analysis of variance showed that these differences among racial groups are also significant but do not account for much variance.

Table 5

Situational Judgment Test (SJT) Scores for Demographic Subgroups

		SJT Total Score		
	N	Mean	SD	R^2
Male	873-867	1.35	.60	
Female	105-109	1.55	.58	.01 *
Black	316-324	1.19	.61	
Hispanic	61-63	1.43	.55	
White	550-557	1.48	.58	.05 *
Other[1]	51-52	1.29	.58	

* Adjusted for shrinkage; significant at the .01 level.

[1] This category was not included in estimating the variance accounted for by race because it is likely that a variety of racial groups are actually included in this one category.

Table 6 shows the mean SJT Total Score for soldiers in combat and non-combat MOS. The average SJT score for soldiers in combat MOS (11B, 13B, and 19E/K) is about a quarter of a standard deviation lower than that for soldiers in the other five MOS. This difference is significant but accounts for very little variance. Table 6 also shows the mean SJT scores for each of the nine different MOS. The MOS with the highest mean scores are 95B and 19E/K, and the MOS with the lowest mean scores include 13B and 88M. Analysis of variance showed that these differences are also significant, and they account for more variance than does the combat/non-combat difference. These differences can be at least partly explained by differences in general cognitive ability. Different MOS have different selection standards, and Table 6 shows the mean Armed

Table 6

Combat/Non-Combat and MOS Differences in Situational Judgment Test (SJT) Scores

		SJT Total Score[1]			AFQT
	N	Mean	SD	R^2	Mean
Combat MOS[2]	278-309	1.27	.61		49.55
Non-Combat MOS[3]	625-687	1.42	.59	.01 *	51.98
MOS:					
11B	116-121	1.25	.62		53.05
13B	128-147	1.21	.63		46.47
19E/K	34-41	1.53	.47		49.24
31C	91-94	1.43	.62		55.57
63B	98-107	1.31	.56		44.13
71L	98-109	1.42	.65		52.13
88M	132-141	1.25	.57		42.25
91A/B	89-100	1.40	.54		57.27
95B	117-136	1.69	.53	.05 *	62.61

* Adjusted for shrinkage; significant at the .01 level.

[1] The rank order correlation between mean SJT Total Score and mean AFQT across the nine MOS is .60.

[2] 11B, 13B, and 19E/K

[3] 31C, 63B, 71L, 88M, 91A/B, and 95B

Forces Qualification Test (AFQT) scores for soldiers in our sample from the various MOS. AFQT scores can be viewed as measures of general cognitive ability (Murphy, 1984). Table 6 also provides the rank order correlation, across MOS, between mean AFQT scores and mean SJT scores.[1] It appears that differences in cognitive ability can account for at least some of the differences in SJT scores across MOS.

---

1. The rank order correlation was used because only nine observations are included in this correlation so a nonparametric test was appropriate.

Conclusions Concerning the Basic Analyses

In general the basic analysis results for the SJT data from the CVII sample are very encouraging. Results show that SJT responses from this sample are adequately spread across the response alternatives and can be used to compute several reliable scores. The M-L Effectiveness score (i.e., SJT Total Score) is quite reliable and appears to provide a good summary of the information contained in all five basic SJT scores. Race and sex differences in SJT scores are small -- less the half a standard deviation -- and account for a very small amount of the total variance in scores. MOS differences are also small and are probably at least partly due to MOS differences in general cognitive ability. The SJT data from the CVII sample have good psychometric qualities and thus provide a good source of information for further explorations of the construct validity of the SJT.

EXPLORATIONS OF THE DIMENSIONALITY OF THE SJT

It is possible that what is measured by the SJT is a single, unidimensional construct that might be labeled something like "knowledge of effective supervisory practices" or "effectiveness of supervisory judgment." However, it is also possible that what is measured by the SJT is actually several different constructs. The SJT item stems describe a wide variety of supervisory situations, and the response alternatives describe a variety of different types of supervisory behavior (e.g., counseling, disciplining, planning and organizing). Thus, it is conceivable that the SJT actually measures several relatively distinct sub-constructs. If distinct sub-constructs can be identified, they could provide a better understanding of what is measured by the SJT. These sub-constructs might also provide the basis for developing SJT subscores. If sub-constructs can not be identified, this would provide at least some support for the notion that the SJT is a unidimensional test. Thus, a thorough investigation of the dimensionality of the SJT was conducted.

Two general approaches were taken to explore the dimensionality of the SJT: empirical and rational. Dimensionality was explored empirically using factor analysis. Several different rational approaches were taken, and each involved categorizing the SJT items and/or response alternatives according to their content. Most of these categorizations were aimed at rationally identifying sub-constructs, but they were also expected to contribute to a more systematic understanding of the content of the SJT. The SJT response alternatives were first categorized according to the supervisory tasks or behaviors involved (e.g., counseling). Additional dimensions along which the SJT items and response alternatives might vary were identified based on a review of leadership and supervision literature, particularly research involving taxonomies of supervisory behavior. Finally, a thorough content analysis of the SJT items and response alternatives along with their effectiveness values was conducted in order to identify other promising dimensions. Each of these approaches is described in more detail below.

Factor Analyses

The item-level scores for each of the three most promising scoring procedures (M-Effectiveness, L-Effectiveness, and M-L Effectiveness) were intercorrelated and factor analyzed using principal factor analysis. From 2 to 5 factors were extracted for each scoring procedure and rotated to a varimax solution. The results did not reveal any clearly defined dimensions and were for the most part uninterpretable. Some partially identifiable factors emerged in a few of these analyses that involved (1) disciplining when appropriate, (2) avoiding disciplining when inappropriate, and (3) assigning work tasks effectively. However, the content of these factors was not very distinct.

Task-Based Content Analysis

Ratings of the extent to which SJT response alternatives involve various supervisory tasks or behaviors were obtained in order to determine the extent to which the SJT taps the various aspects of supervision identified in the earlier job analysis (Campbell, 1989). This was done by first identifying a set of dimensions of supervisory behavior relevant to the SJT. Four

researchers independently content analyzed the SJT response alternatives, developed category systems based on the supervisory tasks or behaviors involved, and sorted the response alternatives into these categories. The four different category systems were rationally collapsed into a single system with ten categories. Table 7 provides a list of these ten categories and a definition of each. Next, nine researchers were asked to rate the extent to which each SJT response alternative fit into each category. Each response alternative was assigned a total of ten points to be divided among the ten categories (in order to ensure that some response alternatives were not over represented in the final category system). These raters were told to assign the points for response alternatives that didn't fit into any of these categories to a miscellaneous category. The interrater reliability of these ratings was estimated for each of the ten categories.

Table 8 shows that the ratings of the supervisory tasks or behaviors involved in the SJT response alternatives were very reliable. For each of the ten categories, this table shows the interrater reliability of the mean number of points assigned to each response alternative (across all nine raters), and these reliabilities range from .82 to .99 with a median of about .94. Reliabilities were particularly high for Referring and for Interacting Assertively with Superiors. Reliabilities were lowest for Reasoning with Soldiers and for Communicating with Subordinates.

The extent to which the SJT measures performance related to each of these ten categories was assessed by computing the mean, across all response alternatives, of the mean number of points the nine raters assigned to each category. The second column of Table 8 presents these overall values. These means are highest for Disciplining and Gathering Information/Monitoring and lowest for Referring, Giving Orders, and Reasoning with Soldiers, but all ten of the categories appear to be adequately represented. Table 9 shows approximately how these dimensions correspond to the dimensions identified in the second-tour job analysis (Campbell, 1989). Two of the dimensions identified in the job analysis, Acting as a Model and Training Subordinates, do not correspond to any of the SJT task categories. For Acting as a Model, this is probably due to the fact that the SJT is a maximal performance measure (i.e., a test), and by its very nature acting as a model is probably better tapped by measures of typical performance (e.g., performance ratings).

These task-based category ratings were also used, in combination with the effectiveness values from the USASMA experts, to determine whether response alternatives involving some supervisory behaviors tend to be more effective than those involving others. The mean effectiveness of each of the ten SJT task-based dimensions across all of the SJT response alternatives was computed using the formula presented on Table 10. For each response alternative, the mean effectiveness rating from the experts was weighted by the extent to which that response alternative was judged to tap a particular dimension. These weighted effectiveness ratings were then added together for each dimension. Finally, in order to place all of the dimensions on the same metric, these dimension scores were divided by the sum (across all response alternatives) of the extent to which the SJT response alternatives tap the relevant dimension. This resulted in effectiveness scores for each of the ten dimensions that are on the same one to seven scale as the effectiveness ratings from the experts,

Table 7

Situational Judgment Test (SJT) Task-Based Categories with Definitions

1. Referring. Refer subordinates to a counseling or help program (e.g. financial counseling, a dietitian, the education center, formal counseling) in response to personal or performance problems.

2. Interacting assertively with superiors. Work assertively with individuals at a higher level in the chain of command, for example to stick up for subordinates' rights, obtain appropriate rewards and punishments for subordinates, or solve subordinates' problems.

3. Counseling. Conduct formal or informal counseling with subordinates concerning performance or personal problems. This includes disciplinary counseling as well as counseling meant to encourage subordinates, help them solve problems, etc.

4. Encouraging. Provide encouragement to subordinates by acknowledging or rewarding good performance or exemplary behavior, also by providing encouragement and support in response to their problems.

5. Disciplining. Discourage inappropriate behaviors or inadequate performance by taking disciplinary actions (e.g. Articles 15, formal counseling statements, additional duty), by warning that disciplinary action may be taken in the future, or by reporting the problem to superiors.

6. Gathering information/monitoring. Gather the information necessary to strategically assign tasks or to take action in response to problems (e.g. poor performance). Monitor subordinates' performance or other behaviors.

7. Reasoning with soldiers. Ensure that subordinates perform assigned tasks and duties by reasoning with them, for example, explaining why the work must be done, providing an incentive, or otherwise persuading them.

8. Giving orders. Give soldiers direct orders, for example orders to perform tasks, activities, or missions.

9. Assigning tasks. Strategically assign tasks in a manner that will best accomplish the mission, address subordinate problems (e.g. performance, personal, or interpersonal problems), or provide developmental opportunities.

10. Communicating with subordinates. Provide subordinates with needed information or advice; keep subordinates informed. This includes communicating specific performance expectations, clarifying tasks or missions, or telling subordinates about opportunities that are available to them.

Table 8

Interrater Reliabilities for SJT Task-Based Category Ratings, Mean Rating for Each Category, and the Mean Effectiveness of Each Category

	Interrater Reliability[1]	Mean Rating Across All Response Alternatives	Mean Effectiveness
1. Referring	.99	.54	4.28
2. Interacting assertively with superiors	.97	1.13	4.44
3. Counseling	.93	1.04	4.49
4. Encouraging	.92	.90	4.10
5. Disciplining	.96	1.43	3.55
6. Gathering information/monitoring	.96	1.46	4.70
7. Reasoning with soldiers	.82	.61	3.87
8. Giving orders	.92	.66	3.89
9. Assigning tasks	.94	1.18	3.89
10. Communicating with subordinates	.86	.85	4.06

[1] Interrater reliabilities are for the mean across nine raters.

Table 9

Correspondence of SJT Task-Based Categories with Job Analysis Dimensions

Dimensions Identified in the Second Tour Job Analysis	Dimensions Identified in the SJT Categorization
Planning Operations	Assigning Tasks
Directing/Leading Teams	Giving Orders
Monitoring/Inspecting	Gathering Information/Monitoring
Individual Leadership	Encouraging
	Reasoning with Soldiers
Acting as a Model	
Counseling	Counseling/Disciplining/Referring
Communicating with Subordinates, Peers and Supervisors	Communicating with Subordinates/Interacting Assertively with Superiors
Training Subordinates	
Personnel Administration	

Table 10

Computation of the Mean Effectiveness of Each SJT Task-Based Category

$$\text{Mean Effectiveness of Each Category} = \frac{\sum_{i=1}^{143} (\text{resp. alt. eff.}^{1})\ (\text{resp. alt. dimension rating}^{2})}{\sum_{i=1}^{143} (\text{resp. alt. dimension rating}^{2})}$$

[1] This is the effectiveness rating from USASMA experts.

[2] For each of the ten dimensions, this is the mean dimension rating across all nine raters.

and these scores are shown in the third column on Table 8. This column shows that response alternatives involving Gathering Information/Monitoring and Counseling tend to be more effective, and those involving Disciplining and Reasoning with Soldiers tend to be less effective.

During the development of the SJT, an effort was made to develop response alternatives involving all the different types of supervisory behaviors identified in the job analysis and also to develop items for which a variety of different of types of supervisory behavior would be the most effective. The results presented in Table 8 indicate that this effort was quite successful. However, because development of the SJT focused on representing all these dimensions as opposed to reflecting the importance of the various dimensions for the job, these results should not be interpreted as reflecting the importance of each of the ten dimensions for the second-tour NCO job.

Dimensions Based on SJT Content Analysis and Relevant Literature

Development and Administration of the Rating Task

A review of the most relevant literature on supervision and leadership and a thorough content analysis of the SJT revealed several additional dimensions along which SJT items and response alternatives could be seen as differing from each other in important ways. After these dimensions were identified and defined, five researchers rated the SJT items and response alternatives on each of these additional dimensions. The procedures used to identify these dimensions and the ratings that were collected for each are described below.

Relevance of Special Training. Most research to date has used situational judgment tests as predictors of job performance rather than as criterion

measures of job performance. As discussed previously, it is likely that the supervisory knowledge or skill assessed by a test such as the SJT has aspects both of an aptitude and of achievement. In addition, it is possible that some SJT items are more clearly measures of achievement (i.e., knowledge that is obtained as a supervisor in the Army) while other items are more like aptitude measures. These latter items are those that could be answered correctly based on general knowledge concerning interpersonal situations or general life experiences, even though these items also reflect correct supervisory practices in the Army. Thus, a rating scale was developed to ascertain the extent to which identifying the more effective responses for each SJT item appears to require special training or knowledge (e.g., familiarity with military supervisory procedures). The rating scale used to collect these ratings is included in Appendix A (labeled Rating Category E).

Sources of Power. One line of leadership research that is particularly relevant to the SJT response alternatives concerns the sources of power and influence tactics that are used by managers and supervisors (e.g., French & Raven, 1959; Podsakoff & Schriesheim, 1985; Kipnis, Schmidt, & Wilkinson, 1980). Power refers to an agent's capacity to influence a target person's behavior, and many researchers have developed taxonomies of the various sources of power or influence tactics that managers or supervisors use to achieve their objectives. The behaviors described in the SJT response alternatives can be seen as attempts to use sources of power or influence tactics to achieve certain objectives (e.g., improved subordinate performance). Taxonomies involving sources of power appeared better suited for categorizing the behaviors described in SJT response alternatives than those involving influence tactics, perhaps because many of the SJT items involve relatively long term rather than immediate influence objectives. French and Raven (1959) developed what is probably the most widely cited taxonomy of power, and this taxonomy has been used by many researchers to study the implications of power for supervisory or managerial effectiveness. Yukl and Falbe (1991) expanded French and Ravens five-factor taxonomy to include three additional sources of power, and this expanded taxonomy was used in the present research. However, two of the sources of power in Yukl and Falbe's taxonomy -- referent power (i.e., likability) and charisma -- appear to be characteristics of a person and not applicable for rating behaviors so these two sources of power were excluded. For each SJT response alternative, ratings were collected concerning which of the remaining six sources of power (if any) was being used in that response alternative in an attempt to influence the target person's behavior. If more than one source of power was being used in a given response alternative, raters were asked to list them in the order of importance. These six sources of power are defined under Rating Category H in Appendix A.

Characteristics of the Situations. Some of the research on power and influence tactics has explored the ways in which aspects of the situation (e.g., crisis vs. non-crisis; involving subordinates, peers, or supervisors; the objective or goal) affect the frequency with which various sources of power or influence tactics are used and how these situational characteristics interact with the sources of power used to affect leadership effectiveness (e.g., Yukl & Falbe, 1990; Mulder, de Jong, Koppelaar, & Verhage, 1986). Accordingly, for each SJT item stem, ratings were collected concerning the direction (upward or downward) of the influence attempt, the objective of the influence attempt,

and (where the relevant information was available) the performance of the target person and this person's typical level of responsibility or maturity. If an SJT situation involved more than one influence objective, raters were asked to list them all in the order of importance. These ratings were used to systematically explore whether aspects of the situation impact on the sources of power used (i.e., chosen in SJT response alternatives) and the effectiveness of the various sources of power. The rating scales used to make these four ratings for each of the SJT item stems are presented as Rating Categories A through D in Appendix A.

Item Types Based on Content Analysis. Finally, because SJT items appear to be more than simply the sum of their parts (i.e., the item stems and the response alternatives), a content analysis was conducted that took into account the content of the item stems, the content of the response alternatives, and the effectiveness of the various response alternatives. The goal of this content analysis was to identify what each SJT item was "getting at" or measuring. For example, an SJT item stem might describe a subordinate who is performing poorly, the more effective response alternatives might involve giving that subordinate a second chance, and the less effective response alternatives might involve disciplining harshly. This item could be seen as tapping the ability to identify situations in which it is most effective to "avoid inappropriately harsh discipline." The first author conducted a thorough content analysis of the SJT items and identified eleven such content-based "item types" that appeared to have potential for identifying relatively homogeneous subsets of SJT items. Ratings were obtained concerning which, if any, of these types captured the essence of each SJT item. Where more than one item type applied, raters were asked to list them in the order of importance.The eleven items types are listed under Rating Category F in Appendix A.

Development of SJT Content Analysis-Based Subscales

Because both the item type ratings and the ratings of the relevance of special training took into account the content of SJT item stems, the content of the response alternatives, and the effectiveness of the response alternatives, these ratings were particularly good candidates for the development of SJT subscales. A set of subscales was developed based on each of these sets of ratings.

Items rated as having the same item type could be seen as measuring or "getting at" the same thing. Therefore, the SJT items were categorized according to the item type involved. Items were grouped into an item type if at least three of the five raters had indicated that it was either the primary or the secondary type for that item. Based on this decision rule, several SJT items were assigned to more than one item type category. In addition, none of the SJT items were assigned to the categories labeled "avoid inappropriately harsh discipline," "clarify performance standards," or "resist being taken in by subordinates' stories," so these categories were dropped. Only a few items were assigned to the item type labeled "searching for underlying personal problems," so these items were assigned to another, similar item type labeled "searching for underlying reasons for problems." Similarly, there were relatively few items that involved "acknowledging or emphasizing the positive," "providing subordinates with needed support and/or encouragement," or

"ensuring that subordinates obtain appropriate rewards," so these were collapsed to form a single item type called "providing support."

This categorization of items into item types was used to form a preliminary set of item type subscales, and scores on these subscales were computed by averaging the item-level M-L Effectiveness scores for the items assigned to each subscale. Because some items had been assigned to more than one item type, some of the subscales contained overlapping items. In an effort to make the subscales more independent, item-total correlations were computed for each of these overlapping items for each of the subscales to which they had been assigned. Items were then dropped from the subscales with which they had lower item-total correlations, so that each item was included on only one item type subscale. Scores were then computed for these revised subscales by averaging the M-L effectiveness scores for each of the items on the subscale.

Ratings of the relevance of special training were also used to develop SJT subscales. First, the interrater reliability of these ratings was evaluated. Based on the conjecture that familiarity with the military and particularly with supervision in the military would aid in making these ratings, the interrater reliability analyses were conducted twice: once including all five raters and once including only the three raters who had the most experience working with the military. Results showed that the interrater reliability was in fact higher for the three raters who had more experience with the military, so only the ratings made by these three raters were used in the development of the special training subscales. The interrater reliability across all 35 SJT items for the mean rating (across the three raters) of the relevance of special training was .63.

The mean rating of the relevance of special training was computed across these three raters for each of the 35 SJT items. For three of the SJT items, the standard deviation of these ratings across the three raters was 1.5 or greater, suggesting that there was a great deal of disagreement concerning the importance of special training for these items. Accordingly, these three items were not included in any special training subscales. The remaining items were then divided into three groups: those rated high, those rated average, and those rated low in terms of how likely it is that special training would be required to identify the more effective response alternatives. Thirteen items had mean ratings between 2 and 2.33 on a five point scale (where "1" indicates that an item definitely requires special training or knowledge and "5" indicates that an item clearly doesn't require special training or knowledge). Eleven of the remaining items had mean ratings between 2.67 and 3.33 on this same five point scale and eight had mean ratings between 4.00 and 4.67. Eight items were randomly selected from each of these three groups of items, so that the resulting scales were equal in length. Scores were then computed for these three special training subscales by averaging the item-level M-L effectiveness scores of the eight items assigned to each subscale. The scale including the eight items with the lowest mean ratings was labeled Training Needed, the scale including the items in the middle group was labeled Training May or May Not be Needed, and the scale including the items with the highest mean ratings was labeled Training Not Needed. These labels are only approximations. Even within each of these groups the items vary somewhat in terms of the importance of special training.

Evaluation of all of the content-based subscales -- those based on the relevance of special training and those based on item types -- involved computing correlations of scores on these subscales with scores on other job performance measures and selected temperament and cognitive ability measures. These correlations are presented in the following section of this report that deals with the relationships of SJT scores with scores on other measures in general.

Relationships Between Sources of Power and Effectiveness

Frequency and Mean Effectiveness of the Various Sources of Power. For each SJT response alternative, frequency counts were conducted to determine how many raters indicated that each source of power was the primary source of power being used in that alternative and how many indicated that each source of power was the secondary source of power. Response alternatives were then assigned to one of the six sources of power if at least three raters had indicated that it was either the primary or the secondary source of power. Those few response alternatives that could be assigned to more than one source of power based on this decision rule were assigned the source of power that was chosen as primary by the most raters. Nineteen of the 143 SJT response alternatives could not be assigned to any of the six sources of power, either because the raters indicated that the source of power was not clear or because there was too much disagreement among the raters concerning the source of power that was being used.

Counts were then made of the number of SJT response alternatives that had been assigned each of the six sources of power, and these frequencies are presented on the right side of Table 11. The percentage of all SJT response alternatives that each of these numbers represent are presented as well. The sources of power that are used most frequently in SJT response alternatives are legitimate, coercive, and information power. Expert power appears the least frequently; only four response alternatives involve the use of expert power. The frequency with which the SJT response alternatives involve the various sources of power is interesting in its own right, but the generalizability of these data for making inferences concerning how frequently these sources of power are actually used is somewhat limited. As discussed previously, development of the SJT response alternatives did involve asking NCOs similar to the target sample what they would do to respond effectively in each situation. However, these responses were then content analyzed and collapsed, and some were dropped because they didn't differentiate between the target sample of NCOs and the expert group. It is thus difficult to say exactly how the frequency with which sources of power occur in the SJT response alternatives relates to the frequency with which they are used in actual supervisory situations in the Army.

The mean effectiveness of SJT response alternatives involving each source of power was also computed, and these are presented in the last column of Table 11. In general, expert and information power appear to be the most effective. However, because there are only four alternatives that involve expert power, any conclusions about this type of power are extremely tentative. The

Table 11

Frequency and Mean Effectiveness of SJT Response Alternatives for Each Direction of Influence and Source of Power

Source of Power	Direction Downward			Direction Upward			Overall		
	Freq.	%	Mean Eff.	Freq.	%	Mean Eff.	Freq.	%	Mean Eff.
Legitimate	37	30%	4.21	1	5%	3.56	38	27%	4.19 (1.25)
Reward	9	7%	3.87	0	0%	--	9	6%	3.87 (1.24)
Coercive	27	22%	3.29	3	15%	4.84	30	21%	3.44 (1.40)
Information	21	17%	5.12	10	50%	4.62	31	22%	4.95 (1.27)
Expert	4	3%	5.17	0	0	--	4	3%	5.17 (0.91)
Persuasive	9	7%	4.27	3	15%	5.06	12	8%	4.47 (1.48)
Not clear or not relevant	16	13%	--	3	15%	--	19	13%	-- --
Total	123	100%	4.07	20	100%	4.55	143	100%	4.14 (1.44)

standard deviation of the effectiveness values for response alternatives involving each source of power are presented in parentheses after the means. Based on the size of these standard deviations, it appears that the distributions of effectiveness values for response alternatives involving each source of power overlap a great deal. However, two-tailed t-tests reveal that the information source is significantly more effective than both the coercive source ($t = 4.42$, $p < .001$) and the legitimate source ($t = 2.50$, $p < .02$). The legitimate source of power is significantly more effective than the coercive source ($t = 2.31$, $p < .03$).

Interactions Between Situation Characteristics and Sources of Power. The ratings of characteristics of the situations described in SJT item stems were used to explore whether certain aspects of these situations affect the frequency with which various sources of power appear in SJT response alternatives or the effectiveness of these sources of power when they appear. Ratings of the immediate performance of the target person and ratings of the typical maturity or responsibility level of the target person were not included in the present analyses for two reasons. First, for about one-third of the SJT items most of the raters indicated that the item stem did not provide this information. Second, there was more disagreement among the raters concerning these two ratings than there was for any of the other ratings.

There was very good agreement among the raters concerning the direction of the influence attempt. All five of the raters agreed on the direction of the influence attempt for almost all of the items, and at least four of the five raters agreed for the remaining items. Based on these ratings, 30 of the SJT items were identified as involving downward influence attempts and the five remaining items as involving upward influence attempts. Frequency counts were then conducted to determine whether the percentage of SJT response alternatives involving each source of power differs in items involving upward versus downward influence attempts.

Table 11 shows that half of the response alternatives for items that involve upward influence attempts describe the use of information power. However, relatively few SJT items involve upward influence attempts, so all of these frequencies are low. The mean effectiveness of response alternatives involving each source of power was computed separately for upward and for downward influence attempts. Results are included on Table 11, and they show a tendency for response alternatives involving persuasive power to be more effective in upward than in downward interactions, but this difference is not significant ($t = .79$, $p > .05$).

SJT items were also categorized according to the objective of the influence attempt involved. Each item was categorized as involving a particular influence objective if at least three of the five raters indicated that it was either the primary or the secondary influence objective. A few items could be assigned to more than one influence objective based on this decision rule, and these were assigned the influence objective that was chosen as primary by the most raters. Two of the 35 SJT items could not be categorized into any of the seven influence objectives and were not include in the present analyses. In addition, only one item had the objective of "providing subordinates with

encouragement and support," so this influence objective and this item were also excluded from the present analyses.

Table 12 shows the number of SJT response alternatives involving each source of power for each of the remaining six influence objectives. These results make a great deal of sense. Coercive power occurs almost exclusively in SJT response alternatives for which the objective described in the item stem is to improve subordinates' performance or to deal with a disciplinary problem. Legitimate power is also involved very frequently when the objective is to improve subordinates' performance. Reward power only occurs in those response alternatives where the objective is to assign tasks or reward performance. Where the objective is to obtain a change in plans, response alternatives almost all involve either information or persuasive power. The mean effectiveness levels of each of these sources of power for each influence objective also show some interesting patterns, but caution should be used in interpreting these means because many of them are based on just a few response alternatives. Response alternatives involving persuasive power appear to be much more effective in those items where the objective is to obtain a change in plans than they are in items involving disciplinary problems ($t = 4.09$, $p < .01$). Those involving information power appear somewhat more effective where the objective is to improve subordinates' performance or to deal with a disciplinary problem than where the objective is to obtain a change in plans, but these differences are not significant ($t = 1.09$, $p > .05$; $t = 1.14$, $p > .05$).

Conclusions Concerning the Dimensionality of the SJT

The item-level factor analyses of the SJT were not very informative. Perhaps this is at least partly due to the multidimensionality of individual SJT items. This hypothesis is supported by the results of the task-based category ratings. Researchers were able to make reliable ratings of the supervisory tasks involved in the SJT response alternatives, and a single SJT item generally involved many different tasks. Many of the individual response alternatives were also rated as involving several different tasks. These task-based ratings also showed that the SJT covers the intended content domain quite well. Conclusions concerning the item type and special training subscales are presented in the next section of this report after correlations between these subscales and other measures have been presented. The analyses concerning sources of power used in the SJT response alternatives provide some interesting information about *why* some responses are more effective than others. These analyses also shed some light on the nature of effective supervisory practices in the Army. Overall, responses that involve information power tend to be most effective and those that involve coercive power tend to be least effective. In addition, characteristics of the SJT situations, specifically the direction and objective of the influence attempt, are systematically related to the relative frequency and effectiveness of the various sources of power. Many of these differences in the mean effectiveness of the various sources of power are intuitively appealing but not statistically significant, and this is at least partly due to the small number of response alternatives in some of the analysis cells.

Table 12

Frequency and Mean Effectiveness of SJT Response Alternatives for Each Objective of Influence and Source of Power

Source of Power	Objective of Influence						
	Assign Tasks	Improve Performance	Reward Performance	Solve Pers. Problems	Disciplinary Problems	Obtain Changed Plans	Overall
Legitimate	4.13 (5)	4.03 (13)	3.07 (3)	4.46 (7)	4.56 (5)	--	4.19 (38)
Reward	3.87 (4)	--	3.86 (5)	--	--	--	3.87 (9)
Coercive	--	3.21 (11)	3.78 (1)	--	3.52 (16)	--	3.44 (30)
Information	4.12 (2)	5.26 (6)	4.75 (5)	5.64 (1)	5.21 (7)	4.49 (7)	4.95 (31)
Expert	--	4.90 (1)	5.83 (1)	4.98 (2)	--	--	5.17 (4)
Persuasive	4.60 (1)	4.43 (2)	--	--	2.63 (3)	5.39 (6)	4.47 (12)
Not clear or not relevant	-- (1)	-- (4)	-- (1)	-- (3)	-- (5)	-- (3)	-- (19)
Total	4.21 (13)	3.99 (37)	4.08 (16)	4.38 (13)	3.78 (36)	4.63 (16)	4.14 (143)

RELATIONSHIPS OF SJT SCORES WITH OTHER MEASURES

In judging the validity of the SJT as a criterion measure of job performance, the relationships between SJT scores and scores on other measures were a key source of information. Criterion related validity is an appropriate method for judging the usefulness of tests as a predictor measures, but the appropriate method for determining the validity of tests as criterion measures is not as straightforward. For the SJT we took a construct validation approach, and the wide variety of temperament, cognitive ability, and job performance scores available for the CVII sample provide an excellent opportunity for construct validation.

Additional Measures Available for the CVII Sample

When the CVII soldiers entered the Army, they were administered the Armed Services Vocational Aptitude Battery (ASVAB), which is a battery of aptitude and ability tests. At the time they took the SJT, most of these soldiers also completed a temperament and biodata inventory called the Assessment of Background and Life Experiences (ABLE). In addition, a wide variety of job performance measures were administered to these soldiers concurrently with the SJT. This included work sample or "hands-on" measures of technical performance and a technical job knowledge test. These soldiers also completed a self-report questionnaire concerning administrative information (e.g., awards, disciplinary actions), and participated in three supervisory simulation (i.e., role play) exercises. Finally, supervisor ratings of these soldiers' job performance were collected at that time as well, using behavior-based rating scales (see Campbell & Zook, 1990 for a complete description of available measures). Scores that were available from each of these measures are described below.

Briefly, the Armed Forces Qualification Test (AFQT) composite of the ASVAB, which was used in the present research, consists of four subtest scores (word knowledge and paragraph comprehension, arithmetic reasoning, and mathematics knowledge), all scores standardized before summing. This composite has been reviewed as a reasonably good measure of "g" (Murphy, 1984). The ABLE was originally designed to measure 10 temperament constructs that had demonstrated criterion-related validity in previous research (Hough, Barge, & Kamp, 1985). Factor analyses of ABLE data were later used to develop a set of shortened factor-based ABLE composites (Campbell & Zook, in preparation). Six of these shortened ABLE composites were hypothesized to be related to SJT scores, either directly or indirectly through other measures, and were therefore included in the present analyses. These composites are: Locus of Control, Cooperativeness, Dominance, Dependability, Stress Tolerance, and Work Orientation.

Regarding the performance measures, first a work sample, "hands-on" test was developed to cover the technical, non-supervisory aspects of performance for each of nine jobs in the research (Campbell, Campbell, Rumsey, & Edwards, 1986). For each test, 15 critical tasks were identified and a technical work sample was developed for each of these tasks. Each task had several performance steps that were scored pass or fail. Two proportion-passed scores were derived for each soldier: one score across all of the tasks that were specific

to that soldier's MOS and one score across all of the general soldiering tasks. Technical job knowledge tests were developed for each of the same nine jobs (Campbell et al., 1986). Knowledge items were written toward a job's 15 non-supervisory tasks, identified in the hands-on test development work, and toward 15 additional critical technical tasks for the job. The knowledge tests were multiple choice, and each test contained 150-200 items. For each soldier, two overall job knowledge test scores were developed: the percentage of correct answers on items measuring MOS-specific tasks and the percentage of correct answers on items measuring general soldiering tasks.

Regarding the self-report administrative action measures, questions were developed requesting information on disciplinary actions, awards/commendations, promotions, and training received (Reigelhaupt, Harris, & Sadacca, 1987). Two measures that were developed based on these questions -- promotion rate and number of military supervisory training courses completed -- were expected to be related to SJT scores and therefore included in the present analyses. The promotion rate variable was computed based on self-report information concerning whether and when these soldiers had been recommended for early promotion and based on information available on a computerized database concerning these soldiers' pay grades and their time in service. These soldiers also reported how long they had been in a supervisory position and how frequently they were required to supervise other soldiers.

The supervisory simulations were developed to assess proficiency in three relatively common supervisory situations. The Personal Counseling simulation requires the assessee to counsel a role-playing assessor whose performance and appearance have been declining. The Disciplinary Counseling exercise presents a more serious counseling problem the assessee must deal with, again with a role-playing assessor. The Training simulation requires the assessee to provide guidance to the assessor who role-plays having difficulty with a common technical task. For each exercise, assessors evaluate assessees on 12-20 three-point BARS scales developed to evaluate performance in that exercise. Exercise scores were formed by summing the scale ratings for an exercise, and then a total score for the simulations was derived by adding together the scores for the three exercises (see Campbell, 1991).

Finally, the performance rating scales were appropriate for assessing effectiveness in any second-tour Army job, and accordingly, were referred to as Army-wide scales (see Campbell, 1991). A variant of the behaviorally anchored rating scale development method (BARS: Smith & Kendall, 1963) was employed in preparing 12 BARS scales. In addition, seven dimensions were identified from the task analysis of the NCO supervisor job. All 19 scales (12 BARS and seven task-oriented scales) were administered to the supervisors of the NCOs in the CVII sample. An average of 1.8 supervisory raters were obtained per soldier in the sample. Factor analysis of the correlations between the scales yielded a four-factor solution. The four rating factors are: Leading/Supervising, Technical Knowledge/Proficiency, Personal Discipline, and Physical Fitness/Military Bearing. A composite was then derived for each factor by unit weighting ratings on the scales that loaded substantially on that factor.

Relationships Between SJT Total Score and Scores on Other Measures

In order to determine whether the SJT does, in fact, measure supervisory job knowledge, several hypotheses can be tested concerning the relationships that would be expected between other measures of job performance, scores on temperament and cognitive ability measures, and scores on a supervisory job knowledge test. First, on the criterion side, the performance rating scales were designed to measure typical performance, while a supervisory job knowledge test is best viewed as a measure of maximal performance. Other researchers have obtained relatively low correlations between measures of typical and maximal performance (e.g., Sackett, Zedeck, & Fogli, 1988). Thus, we would expect the correlations of the SJT Total Score with job performance ratings to be only moderate. In addition, the correlation of SJT Total Score with the Leading/Supervising rating composite is expected to be higher than its correlations with the other nonsupervisory rating composites.

Both the SJT and the supervisory simulation exercises were designed to be maximal performance measures in the supervisory part of the job. However, the supervisory simulations were designed to measure only two of the supervisory task categories that were identified in the job analysis, while the SJT appears to tap seven of these nine categories (see Tables 8 and 9). In addition, the SJT focuses on knowing what to do in difficult supervisory situations, while the simulations also measure knowing how to carry these tasks out and skill in doing so. Consequently, only moderate correlations between the supervisory simulations and SJT are expected. Scores on the three supervisory measures (i.e., the Leading/Supervising ratings, the SJT, and the simulations) are also expected to correlate more highly with each other than they do with measures of other, non-supervisory aspects of job performance. However, all three of the measures of supervisory job performance involve different measurement methods, so the error associated with each of these measurement methods may further reduce correlations between these measures.

Temperament and cognitive ability measures provide additional information concerning the construct or constructs measured by the SJT. Certain temperament scales, such as Dominance, would be expected to correlate more highly with a measure of supervisory job knowledge than they do with technical work sample or job knowledge test scores. In addition, SJT scores are expected to correlate at about the same level as the technical job knowledge test scores with cognitive ability.

The degree to which the knowledges and skills assessed by the SJT are learned on the job is central in determining whether it is appropriate as a criterion measure. Information is available concerning how long the soldiers in the CVII sample have been in supervisory positions, how frequently they are required to supervise other soldiers, and the number of supervisory training courses they have attended. If the SJT is a measure of supervisory job knowledge or skill, soldiers who have more experience and training would, on average, obtain higher scores than soldiers with less experience or training. Thus, the mean SJT Total Scores for soldiers with differing levels of experience and training were examined, and correlations of SJT Total Scores with amount of supervisory experience and training were also examined.

Results for Other Criterion Measures

Table 13 presents the correlations of the SJT Total Scores with scores on other job performance measures. As expected, SJT correlations with the supervisory simulations and the Leading/Supervising ratings are significantly different from zero but only moderate in size. Table 13 also shows the correlations between SJT Total Scores, supervisory simulation scores, Leading/Supervising ratings, and selected measures of non-supervisory aspects of job performance. Most of the non-supervisory measures included on this table are measures of technical aspects of job performance. The SJT has its highest

Table 13

Correlations Between Supervisory Performance Measures and Other Job Performance Measures

	SJT Total Score	Supervisory Simulations	Lead./Sup. Rating
SJT Total Score	(.75)		
Supervisory Simulations	.20	(.72)	
Leading/Supervising Rating	.22	.15	(.64)
Job Knowledge - Gen. Soldiering	.39	.19	.17
Job Knowledge - MOS-Specific	.37	.23	.16
Hands On - Gen. Soldiering	.11	.15	.08
Hands On - MOS-Specific	.14	.12	.13
Technical Know./Prof. Rating	.20	.13	.81
Personal Discipline Rating	.19	.07	.67
Phys. Fitness/Mil. Bearing Rating	.10	.10	.60
Promotion Rate	.22	.18	.33

Note. Sample sizes range from 774 to 1020. Correlations larger than about .08 are significantly different from zero at the $p < .01$ level; all correlations are significantly different from zero at the $p < .05$ level. Numbers that appear in parentheses are reliability estimates.

correlations with the technical job knowledge test scores. Both are multiple choice, paper-and-pencil tests, so this correlation is at least partly due to shared method variance. The SJT, the Leading/Supervising ratings, and the supervisory simulations correlate significantly with many of the measures of technical performance, suggesting that the technical and supervisory aspects of second-tour soldiers' job performance overlap to some extent. The correlation between SJT Total Score and the Leading/Supervising ratings is higher than the SJT correlations with the non-supervisory rating composites, but the differences are small and not significantly different from zero for the Technical or Personal Discipline ratings.

Table 13 also shows that promotion rate is moderately correlated with all three measures of supervisory job performance, and this lends additional support for their construct validity. Soldiers who are promoted to higher pay grades are probably given more opportunities to develop their supervisory knowledges and skills, and soldiers who show supervisory potential are probably also more likely to be promoted.

Results for Temperament and Cognitive Ability Measures

Table 14 shows the correlations between scores on the AFQT and the temperament scales from the ABLE and the measures of supervisory job performance. The SJT correlates most highly with the AFQT, and this correlation is probably at least partly due to shared method variance (both are both multiple choice, paper-and-pencil tests). The correlation between the AFQT and the SJT is quite large in light of the fact that these measures were administered several years apart in time. As expected, the correlation between the SJT and the AFQT is at about the same level as those between the job knowledge test scores and the AFQT. The SJT has moderate correlations with Dominance, Dependability, and Work Orientation and is slightly correlated with Locus of Control and Stress Tolerance. All of the measures of supervisory job performance have very similar patterns of correlations with the ABLE scales with two exceptions. Dependability correlates higher with the SJT than it does with the other two supervisory measures, but the difference is only significant for the Leading/Supervising rating ($Z_1^* = 2.21$, $p < .01$)[1]. Work Orientation correlates significantly higher with the Leading/Supervising rating than it does with the SJT ($Z_1^* = 2.27$, $p < .01$) or with the supervisory simulation ($Z_1^* = 1.86$, $p < .05$). Because these correlations are based on about 500 people, a fairly substantial difference between two correlations is required to reach statistical significance (about .10). Finally, it is interesting to note that AFQT scores are correlated with the SJT and the supervisory simulations, but are not significantly correlated with the Leading/Supervising ratings.

Table 14 also shows the relationships between the AFQT and temperament scales and scores on the technical performance measures. The patterns of correlations with the AFQT are similar for supervisory and technical performance.

---

1. Tests comparing elements of correlation matrices were conducted using Z_1^*, as recommended by Steiger (1980). See Steiger (1980) for details concering the exact nature of this test and the rationale for its use.

Table 14

Correlations Between Job Performance Measures and Temperament and Cognitive Ability Measures

	SJT Total Score	Supervisory Simulations	Lead./Sup. Rating	JK Gen. Sold.	JK Specific	HO Gen. Sold.	HO Specific	Technical Rating
AFQT	.31	.16	.06	.37	.38	.13	.15	.10
ABLE:								
Locus of Control	.11	.10	.15	.12	.03	.02	-.03	.14
Cooperativeness	.09	.06	.11	.09	.02	-.01	.03	.08
Dominance	.20	.21	.24	.14	.02	.06	.09	.18
Dependability	.22	.13	.09	.16	.13	-.01	.00	.10
Stress Tolerance	.11	.11	.15	.15	.07	.12	.08	.08
Work Orientation	.20	.17	.31	.16	.09	.01	.03	.27

Note. Sample sizes for the AFQT range from 723 to 893; all correlations are significantly different from zero at the $p < .01$ level. Sample sizes for the ABLE range from 466 to 577; correlations larger than about .11 are significantly different from zero at the $p < .01$ level.

The correlations between Dominance, Dependability, and Work Orientation and the measures of supervisory performance tend to be higher than those between these same temperament scales and the measures of technical performance, and some are significantly higher. For example, the correlation between Dominance and scores on the supervisory simulations is significantly higher than that between Dominance and the technical hands-on test scores ($Z_1^* = 2.12$, $p < .01$)

Results for Supervisory Experience and Training

Table 15 shows the mean SJT Total Scores for soldiers who reported various levels of supervisory training. Soldiers who had attended no supervisory school at all scored significantly lower than those who had attended one or more supervisory schools ($t = 6.75$, $p < .001$), almost a half a standard deviation lower. Some soldiers in this sample had attended more than one supervisory school. The first supervisory school that soldiers typically attend is PLDC, and the next level of supervisory training they typically receive is called BNCOC. Thus, the mean SJT Total Score was computed for soldiers who had attended PLDC only and compared with the mean SJT Total Score for soldiers who had attended both PLDC and BNCOC. Table 15 shows that this latter group of soldiers did score higher than the former, but the difference is small and not significantly different from zero. The correlation between SJT Total Score and the number of supervisory training courses completed (0, 1, or 2) is also presented in Table 15. One potential confound in all of these analyses involving supervisory training is that the opportunity to attend supervisory schools varies, and decisions concerning which soldiers are given the opportunity to attend these schools may be influenced by their effectiveness as soldiers or as supervisors. It is possible that mean SJT score differences were obtained because the more intelligent and more effective soldiers were promoted faster and thus given the opportunity to attend supervisory training. (However, results of the structural modeling analyses presented later in this report suggest a slight tendency for soldiers who score lower on general mental ability to be sent to supervisory training more often.) Regardless of whether these differences are due to differential opportunities or to training in the relevant supervisory knowledges and skills, these mean score differences provide some support for the construct validity of the SJT as a measure of supervisory job knowledge or skill.

Soldiers in the CVII sample were also asked to report how frequently they are required to supervise other soldiers, and mean SJT Total Scores are also reported in Table 15 for subgroups of soldiers identified by their responses to this question. The expected pattern was found, more frequent supervisory responsibilities are associated with higher SJT scores. The correlation between self-reported frequency of supervisory responsibilities (on scale of 1 to 4) and SJT Total Score is relatively low but significantly different from zero. Table 15 also shows the correlation between self-reported time in a supervisory position and SJT Total Score. This correlation is also fairly low but significant. These correlations are somewhat smaller than that between SJT scores and supervisory training.

The relationships of SJT scores with supervisory experience and training are consistent but fairly small in the CVII sample. However, it should be kept

Table 15

Relationships Between Situational Judgment Test (SJT) Scores and Amount of Supervisory Training and Experience

		SJT Total Score		
	N	Mean	Std. Dev.	Corr.[1]
Number of supervisory training courses completed:	955			.20
Attended one or more supervisory schools	593	1.47	.57	
Attended no supervisory school	362	1.20	.63	
Attended PLDC	506	1.46	.57	
Attended PLDC and BNCOC	84	1.53	.58	
How often required to supervise other soldiers:	984			.15
Never	98	1.24	.65	
Sometimes fill in for regular supervisor	334	1.29	.61	
Often fill in for regular supervisor	122	1.38	.63	
Regularly supervise other soldiers	420	1.47	.56	
Time in a supervisory position	809			.14

[1] Correlations of about .07 or greater are significant at the .05 level.

in mind that this sample represents a relatively narrow range of levels of supervisory experience and training. Soldiers in this sample range from having virtually no supervisory experience to having a few months or at most a few years of experience. Similarly, supervisory training for this sample ranges from none at all to one or two low level supervisory courses. Given this narrow range of experience and training, the fact that there is any relationship with SJT scores at all is encouraging.

Data from the SJT developmental work provide some information about SJT responses from groups of soldiers with more widely differing levels of experience. As discussed previously, effectiveness ratings for each response alternative for the 180 candidate SJT situations were collected both from a group of the target NCOs (i.e., beginning supervisors) and a group of very senior NCOs. The senior NCOs had higher levels of agreement concerning the effectiveness of the SJT response alternatives than did the target NCOs. Within each group, the standard deviation of the effectiveness ratings for each response alternative was computed. Across all of the candidate situations (i.e., items), the average standard deviation of effectiveness scale values for the target NCOs was 1.54 whereas for the senior group the mean standard deviation was only 1.39 ($t = 231.91$, $p < .001$). The fact that more experienced NCOs agreed more than did the target sample concerning the effectiveness of various responses to these situations is more pronounced in a count of the number of response alternatives for which the standard deviation of scale values is larger for novices than for experts. For 69% of the response alternatives, across all 180 situations, this standard deviation was larger for the target NCOs than for senior NCOs. Because the 35 situations for the final test were chosen based on high levels of senior NCO and low levels of target NCO agreement, these differences should be substantially larger for the actual test.

Differences between known groups provide some information concerning whether a test measures knowledge or ability, but these comparisons are likely confounded by the fact that in many situations the more experienced group is also a highly selected group. If selection related to ability has occurred, the more experienced group would be expected to score higher on both knowledge and ability tests.

Relationships Between SJT Subscales and Scores on Other Measures

If the SJT subscales based on the relevance of special training and item type ratings actually measure different subconstructs they would be expected to have somewhat different patterns of correlations with the other performance measures and with the measures of temperament and ability. For example, scores on the Training Needed subscale should correlate more highly with amount of supervisory training and experience than do scores on the Training Not Needed subscale. The Training Not Needed subscale, in turn, might be expected to correlate more highly with general cognitive ability (i.e., AFQT scores). For the item type subscales, scores on the Disciplining When Appropriate subscale might correlate more highly with Dependability than do the other subscales, while the Searching for Underlying Reasons subscale might be more strongly related to general cognitive ability.

Table 16 presents the correlations between supervisory training, experience, and job performance measures and the three rationally developed SJT subscores based on ratings of the importance of special training: (1) the mean M-L Effectiveness score across the eight SJT items rated highest in terms of the importance of relevant training (Training Needed); (2) the mean M-L Effectiveness score across eight SJT items rated as average in terms of the importance of relevant military training (Training May or May Not be Needed); and (3) the mean M-L Effectiveness score across the eight SJT items rated lowest in terms of need for relevant training (Training Not Needed). In general, these scores show the expected pattern of correlations with supervisory training, experience, and performance, but differences in correlations across the three SJT subscales are generally small. The Training Needed subscale has a significantly higher correlation with number of supervisory training courses completed than do the other two subscales ($Z_1^* = 1.72$, $p < .05$; $Z_1^* = 1.74$, $p < .05$). The pattern is in the expected direction for the measures of supervisory experience and for the supervisory simulation scores, but the differences are very small and nonsignificant. The Leading/Supervising ratings correlates quite a bit more highly with the Training Needed subscale than it does with the Training Not Needed subscale ($Z_1^* = 2.10$, $p < .01$). In general, these results provide only modest support for the notion that some SJT items are more clearly measures of achievement while others perform more like ability test items.

Table 16 also shows the correlations of these three SJT training related subscales with scores on the AFQT and the ABLE. These correlations are somewhat more difficult to interpret. Two of the ABLE scales -- Dominance and Work Orientation -- have patterns of correlations with these three SJT training related subscales that are similar to the pattern of correlations these subscales have with supervisory training and Leading/Supervising ratings. Perhaps soldiers scoring higher on Dominance and Work Orientation are more likely to obtain the relevant supervisory experience, training, and knowledges, while they are not necessarily more likely than others to have the abilities that can be used to respond effectively to the SJT items that do not require relevant training. This notion is supported by the fact that scores on the AFQT are moderately correlated with scores on the Training Not Needed subscale, but AFQT scores are also correlated at about the same level with scores on the Training Needed subscale. Dependability has a rather curious pattern of correlations with these SJT subscales: correlations between Dependability and scores on the Training May or May Not be Needed subscale are lower than those for the other two subscales, but this difference is only significant for the Training Needed subscale (Z_1^* 1.94, $p < .05$). There is no readily apparent explanation for this result.

Table 17 shows correlations between the SJT subscales based on item type ratings and measures of supervisory performance, experience, and training. When interpreting these correlations, it should be kept in mind that the SJT item type subscales vary widely in the number of items included in each subscale and their internal consistency reliabilities. This information is available at the bottom of Table 17. Correlations of the various item type subscales with Leading/Supervising ratings and supervisory simulation scores are generally similar, which is somewhat surprising in light of the varying

Table 16

Correlations of SJT "Training Needed" Subscales with Each Other and with Scores on Other Measures

	Training Needed	May or May Not be Needed	Training Not Needed	Total Score	Sample Size
Supervisory Simulations	.16	.17	.13	.20	872
Leading/Supervising Rating	.20	.15	.12	.22	797
Supervisory Training	.20	.14	.14	.21	952
Supervisory Time	.12	.11	.10	.14	808
Supervisory Frequency	.12	.12	.10	.15	963
AFQT	.24	.19	.23	.31	841
Dominance	.18	.16	.09	.20	545
Dependability	.23	.14	.19	.22	544
Work Orientation	.20	.15	.10	.20	544
Subscale Intercorrelations:					
Training Needed	(.35)	--	--	.76	984
May or May Not be Needed	.43	(.38)	--	.74	984
Training Not Needed	.44	.43	(.44)	.76	984
Number of Items	8	8	8	35	

Note. Note. Numbers in parentheses are internal consistency reliability estimates. All correlations involving performance or experience are significantly different from zero at the $p < .01$ level. For the ABLE, correlations of about .11 are significant at the $p < .01$ level. All correlations involving the AFQT are significantly different from zero at the $p < .01$ level.

Table 17

Correlations of SJT Item Type Subscales with Each Other and with Scores on Other Measures

	Disciplining as Appropriate	Providing Support	Searching for Reasons	Assigning Tasks	Chain of Command	Total Score	Sample Size
Supervisory Simulations	.16	.17	.14	.17	.05	.20	872
Leading/Supervising Rating	.18	.16	.13	.16	.15	.22	797
Supervisory Training	.20	.15	.13	.13	.11	.21	952
Supervisory Time	.11	.09	.11	.03	.09	.14	808
Supervisory Frequency	.15	.11	.08	.12	.12	.15	963
AFQT	.15	.25	.30	.18	.25	.31	841
Dominance	.18	.16	.09	.14	.10	.20	545
Dependability	.24	.15	.12	.11	.07	.22	544
Work Orientation	.21	.16	.10	.12	.11	.20	544
Subscale Intercorrelations:							
Disciplining as Appropriate	(.44)	--	--	--	--	.67	984
Providing Support	.30	(.67)	--	--	--	.67	984
Searching for Reasons	.31	.55	(.82)	--	--	.82	984
Assigning Tasks	.25	.31	.35	(.56)	--	.56	984
Chain of Command	.26	.43	.48	.30	(.59)	.59	984
Number of Items	7	7	10	4	4	35	

Note. Numbers in parentheses are internal consistency reliabilities. For the performance measures, correlations of about .09 and above are significant at the $p < .01$ level. For the ABLE, correlations of about .11 are significant at the $p < .01$ level. All correlations involving the AFQT are significantly different from zero at the $p < .01$ level.

lengths and reliabilities of these item type scales. The one exception is the lack of a significant correlation between scores on the supervisory simulations and the item type subscale that involves working through the proper chain of command as appropriate. The simulations were designed to measure skill in dealing with subordinates and whereas this Chain of Command subscale is the only item type subscale that does not involve dealing with subordinates, so this difference lends some support to the construct validity of these subscales. Finally, in light of the fact that the Searching for Reasons subscale is the longest and most reliable subscale, it is interesting that scores on this subscale do not correlate more highly with scores on some of these other supervisory measures.

Table 17 also presents the correlations of the SJT item type subscales with AFQT and ABLE scores. As expected, the Searching for Reasons subscale has the highest correlation with AFQT scores, but this correlation is not significantly higher than the correlations of the Providing Support or Chain of Command subscales with AFQT scores. Perhaps this subscale is functioning somewhat like an ability measure (e.g., a measure of problem solving ability). The subscale called Disciplining as Appropriate correlated more highly with Dependability than did any other SJT item type subscale ($Z_1^* 1.75$, $p < .05$). This suggests that the correlation between the SJT total score and Dependability can, to a large extent, be attributed to skill in recognizing when discipline is necessary and appropriate. Perhaps soldiers who are less dependable themselves are more willing to give others a second chance, even when discipline would be a more appropriate and effective response.

Structural Modeling

Rationale and Procedures

The correlational analyses described to this point provide a great deal of information concerning which job performance, experience, temperament, and cognitive ability measures are related to SJT scores. However, there is reason to believe that some of these variables actually affect supervisory job knowledge indirectly, through their relationships with other variables. For example, it seems plausible that the relationship between Dominance and SJT scores is at least partially the result of more dominant individuals obtaining more opportunities to supervise other soldiers and to attend supervisory training. This experience and training, in turn, may be what actually leads to more supervisory knowledge. In addition, supervisory job knowledge might be expected to mediate other relationships, such as the relationship between supervisory training and supervisory skill (i.e., supervisory simulation scores), or between individual differences such as AFQT scores and ratings of supervisory/leadership job performance. If the SJT exhibits the direct and indirect relationships that are expected for supervisory job knowledge, there would be additional support for the construct validity of the SJT. Thus, a series of hypotheses concerning how indirect causal relationships might account for the observed correlations between SJT scores and other variables were developed and tested using the CVII sample data. Four figures are presented later in this section that provide pictorial descriptions of each of these hypotheses.

Hypothesis 1 proposes that the correlation between frequency of supervisory responsibilities (i.e., supervisory experience) and SJT scores can be at least partly accounted for by the relationships of these two variables with supervisory training. In other words, soldiers who have more supervisory responsibilities are more likely to be sent to supervisory training and this training, in turn, leads to higher SJT scores.

Hypothesis 2 postulates that the correlation between certain individual differences measures (AFQT, Dominance, Dependability, and Work Orientation) and SJT scores can be at least partly accounted for by the relationships of all of these variables with supervisory training and experience. In other words, soldiers who score high on certain individual differences measures are more likely to obtain more supervisory training and experience and this in turn leads to higher SJT scores.

Hypothesis 3 proposes that the correlation between certain individual differences measures (AFQT, Dominance, Dependability, and Work Orientation), scores on the supervisory simulation, and ratings of supervision/leadership can be at least partly accounted for by their relationships with SJT scores. In other words, soldiers who score high on certain individual differences measures are likely to obtain more supervisory job knowledge (i.e., higher SJT scores) and this in turn leads to higher scores on the supervisory simulations and higher supervision/leadership ratings.

Hypothesis 4 suggests that the correlation of frequency of supervisory responsibilities and supervisory training with scores on the supervisory simulations and ratings of supervision/leadership can be at least partly accounted for by their relationships with SJT scores. That is, soldiers who obtain more experience and training are likely to obtain more supervisory job knowledge (i.e., higher SJT scores) and this in turn leads to higher scores on the supervisory simulation and higher supervision/leadership ratings.

Structural model analysis provides a tool for exploring the extent to which indirect relationships can account for observed correlations. For each of the four hypotheses, several nested models were developed and compared to determine the extent to which the observed correlations can be accounted for by indirect effects. This was done by first intercorrelating the variables involved in each hypothesis. These correlation matrices, along with reliability estimates for each variable, were then analyzed using the LISREL VI computer program (Joreskog & Sorbom, 1981). Procedures for using the reliabilities to link the observed variables to the latent constructs are identical to those used by Borman, Hanson, Oppler, Pulakos, and White (in preparation). LISREL VI is designed to analyze covariance structural models and is only appropriate for analyzing correlation matrices if the models to be tested are scale invariant. In order to determine whether the use of correlation matrices was appropriate in the present analyses, all analyses were conducted a second time using the variance-covariance matrices, as suggested by Cudeck (1989). Results indicated that correlation matrices are, in fact, appropriate for the models tested, and only the correlational results are presented here.

Reliability estimates used in these analyses are presented on the diagonals of the relevant correlation matrices. For the ABLE composites these

estimates are internal consistency reliabilities (see Campbell & Zook, in press). For the AFQT, the reliability estimate used is the test-retest reliability of the AFQT composite (McCormick, Dunlap, Kennedy, & Jones, 1983). The reliability estimates for time as a supervisor and frequency of supervisory responsibility are rational estimates of the test-retest reliabilities of these measures, taking into account that they were collected via self-report. The internal consistency of the SJT (described earlier) was used as its reliability estimate, and the reliability for the supervisory simulations was estimated by essentially averaging the interrater reliability coefficients across the rating scales for the 79 assessees in the sample that were evaluated by two assessors (see Campbell, 1991). The reliability of the Leading/Supervising rating composite is the intraclass correlation interrater reliability of the mean across the 1.8 supervisor raters per ratee.

The models related to Hypothesis 2 and Hypothesis 3 were each fitted four times, once for each of the four individual differences variables of interest. ABLE scores (i.e., Dominance, Dependability, and Work Orientation) were only available for about 60 percent of the total CVII sample, because only those CVII respondents who completed the other paper-and-pencil measures (i.e., job knowledge tests, SJT, etc.) early were administered the ABLE. It is likely that those who completed these other paper-and-pencil measures early differed systematically from those who did not. In particular, those who finished early are probably higher in general mental ability than those who did not, and their mean AFQT score is in fact significantly higher than that for the total sample ($t = 53.68$, $p < .001$). This could lead to a restriction in the range of AFQT scores in the subsample of respondents with ABLE scores. Accordingly, the total sample was used in all tests involving the AFQT.

For all four hypotheses, LISREL VI was used to obtain estimates of the parameters for each of the models tested. The fit of each of the models to the data was evaluated several ways. The chi-square statistic was computed for each model because of its usefulness in testing the significance of the differences in fit between different nested models (e.g., Mulaik et al., 1989). The probability of these chi-square values was evaluated in two ways, once against the more traditional null hypothesis of exact or perfect fit and once against a less stringent null hypothesis of "close fit" (Browne & Cudeck, in preparation). The null hypothesis of exact fit is invariably false in practical situations and is likely to be rejected when using large samples. The null hypothesis of close fit is an attempt to circumvent these problems. For three of the four hypotheses tested in the present research, several models containing the same set of variables are tested and compared. Because these tests were not independent, it was necessary to take the number of models tested into account in interpreting the probability of the chi-square values. For those hypotheses that involved testing several models, we required a probability of .10 rather than .05 to accept a model as fitting adequately in order to control experiment-wide error. In many cases the same variables were included in models used to test several different hypotheses. However, in these cases each hypothesis was focused on different relationships among the variables, so the tests of the various hypotheses were treated as independent of each other.

For each model, we also computed an estimate of the overall population discrepancy function, which essentially describes the dissimilarity between

the population covariance matrix and the estimated covariance matrix for a particular model that is fitted in the sample. Because these estimates contain a certain amount of error, we present the 90% confidence interval for each of these estimates as recommended by Browne and Cudeck (in preparation). Finally, for each model we computed the Root Mean Square Error of Approximation (RMSEA), which can be interpreted as a measure of the discrepancy per degree of freedom for the model (Browne & Cudeck, in preparation), and the 90% confidence interval for this RMSEA. Browne and Cudeck suggest that a value of .08 or less for the RMSEA can be interpreted as indicating a reasonable error of approximation for a model.

Results Related to Hypothesis 1

For Hypothesis 1, a single model was tested, and this model is presented in Figure 1. Because this model is missing only one possible path or parameter (i.e., has only one degree of freedom), this model essentially tests whether the observed correlations can be accounted for without that one path: the supervisory experience->SJT path. The correlation matrix and reliability estimates used in testing this model are presented in Table 18. The LISREL results and fit indices for this model are presented in Table 19, and these results show that this model is very consistent with the data. The population discrepancy function is .00, and the probability of close fit indicates that the model fits the data very well. The probability of perfect fit is significant at the .05 level, but not at the .01 level. These results indicate that experience->SJT path is not needed, and in the tests of the remaining hypotheses that include these same variables this path is not included.

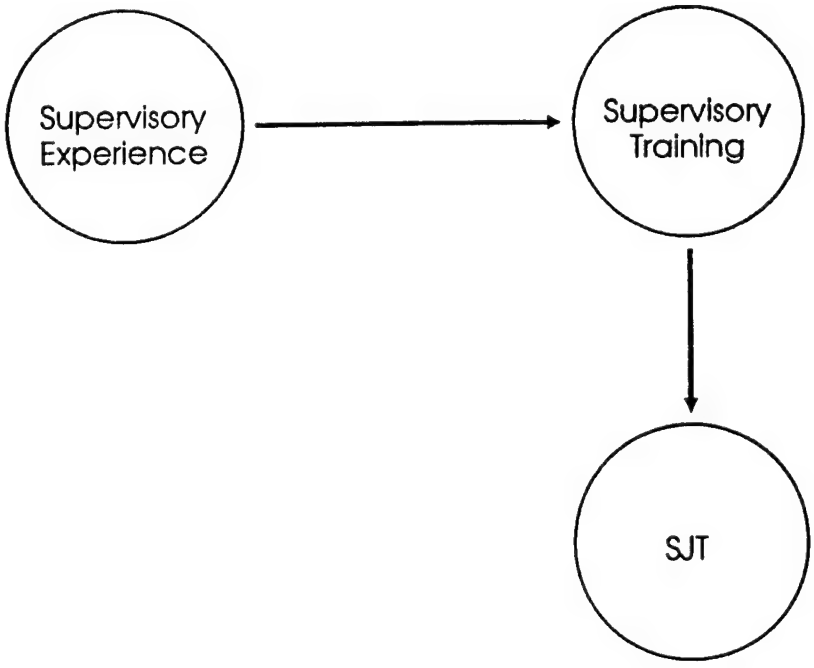

Figure 1. Hypothesis 1.

Table 18

Correlation Matrix Used in Testing Hypothesis 1

	Supervisory Experience	Supervisory Training	SJT
Supervisory Experience	(.85)		
Supervisory Training	.38	(.85)	
SJT	.15	.20	(.75)

Note. Sample size is 933.

Table 19

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Hypothesis 1

Experience->Training Path	.45
Training->SJT Path	.25
Chi-square (df)	4.21 (1)
Probability of Perfect Fit	p=.04
Probability of Close Fit	p=.30
Population Discrepancy (90% Confidence Interval)	.00 (.00-.02)
RMSEA (90% Confidence Interval)	.06 (.01-.12)

Note. Sample size is 933.

Results Related to Hypothesis 2

Hypothesis 2 was tested four times, once for each of the individual differences measures of interest: AFQT, Dominance, Dependability, and Work Orientation. For each individual differences measure, three models were tested, and Figure 2 summarizes the information included these three models. The solid lines represent paths that were included in all three models, and the dashed lines represent paths that were included in only a subset of these models. The first model, Model A, did not include the direct paths from the individual differences measure to supervisory training or to SJT scores (i.e., it included only the paths represented by solid lines). Model B was identical to Model A, except that the path from the individual differences measure to supervisory training was also included. Model C included all of the paths shown in Figure 2. Because Model A is nested within Model B and Model B is nested within Model C, comparisons between these models can address the importance of the direct effects of individual differences on supervisory training and on SJT scores.

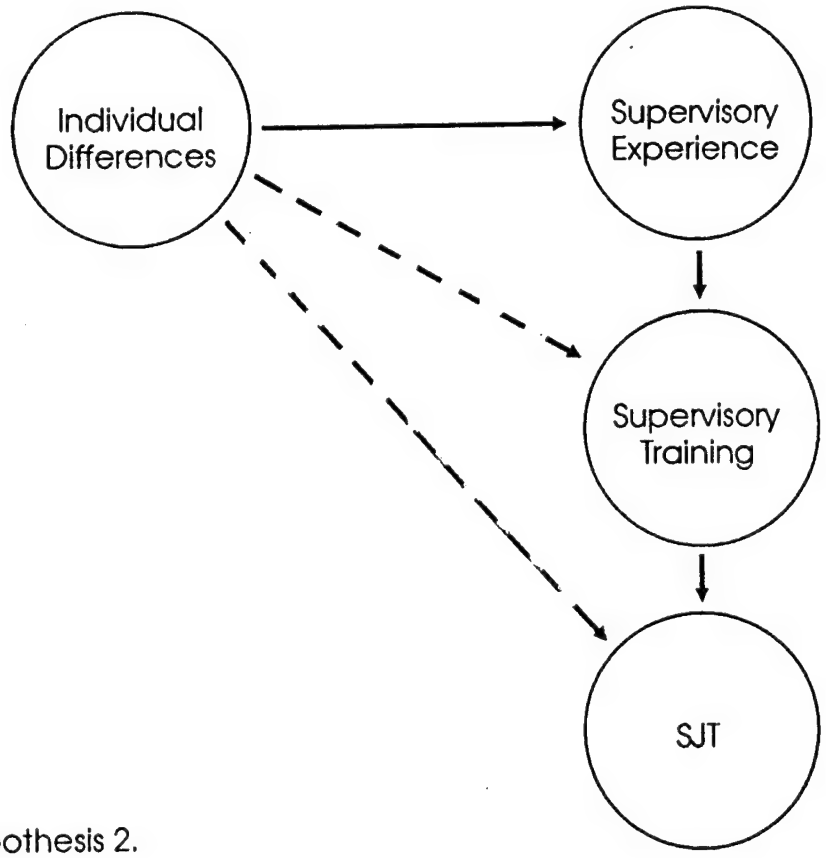

Figure 2. Hypothesis 2.

Table 20 shows the correlation matrix and reliability estimates used to test Hypothesis 2 for the AFQT. Table 21 presents the LISREL results and fit indices for the three models tested. These results provide no support for the hypothesis that the correlation between the AFQT and SJT scores can be accounted for by supervisory training and experience. The models without a direct path from the AFQT to the SJT do not fit the data very well, and

Table 20

Correlation Matrix Used in Testing Hypothesis 2 for AFQT Scores

	AFQT	Supervisory Experience	Supervisory Training	SJT
AFQT	(.93)			
Supervisory Experience	.11	(.85)		
Supervisory Training	-.04	.38	(.85)	
SJT	.30	.17	.14	(.75)

Note. Sample size is 622.

conclusions are the same whether the tests of perfect or close fit are used. In addition, the chi-square statistics for Models A and B are significantly higher than the chi-square for Model C. The AFQT->SJT path in Model C is significantly different from zero, and it is moderately large (.37). Although neither Model A nor Model B fits the data very well, the RMSEA results suggest that when the number of degrees of freedom are taken into account the fit for Model B is actually worse than that for Model A. This indicates that the direct path from AFQT to training is not necessary. It is somewhat surprising that, in both Models B and C, this AFQT->supervisory training path is actually negative. This suggests that if supervisory experience is held constant there is some tendency for lower ability soldiers to be sent to supervisory training. However, in both of these models the AFQT->supervisory training path is very small, and these results should not be over interpreted.

Table 22 shows the correlation matrix and reliability estimates used to test both Hypothesis 2 and Hypothesis 3 for all three temperament variables: Dominance, Dependability, and Work Orientation. LISREL results and fit indices related to Hypothesis 2 for Dominance are presented in Table 23. These results show that Model A fits the data fairly well, with a chi-square of 15.95 and a population discrepancy of only .03, but even the test of close fit rejects the model. The fit for Model B is quite a bit better, and the fit for Model C is nearly perfect. Taken together, these results indicate that Dominance has a large direct effect on supervisory experience, and that its indirect effect through experience can account for its relationships with supervisory training and SJT scores fairly well. Howêver, including the Dominance->training path and including the Dominance->SJT path both improve the fit of the model to the data, suggesting that Dominance has at least some direct effect on amount of supervisory training received and on SJT scores that can't be accounted for by supervisory experience.

Table 21

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 2 for AFQT

	Model A	Model B	Model C
AFQT>Experience Path	.11	.12	.12
AFQT->Training Path	--	-.09	-.10
AFQT->SJT Path	--	--	.37
Experience->Training Path	.45	.46	.46
Training->SJT Path	.18	.18	.19
Chi-square (df)	71.29 (3)	66.55 (2)	4.29 (1)
Probability of Perfect Fit	$p < .001$	$p < .001$	$p = .04$
Probability of Close Fit	$p < .001$	$P < .001$	$p = .21$
Population Discrepancy (90% Confidence Interval)	.11 (.07-.16)	.10 (.07-.15)	.01 (.00-.02)
RMSEA (90% Confidence Interval)	.19 (.15-.23)	.23 (.18-.28)	.07 (.01-.15)

Note. Sample size is 622.

Table 22

Correlation Matrix Used in Testing Models Related to Hypotheses 2 and 3 for Dominance, Dependability, and Work Orientation

	Dominance	Dependability	Work Orientation	Supervisory Experience	Supervisory Training	SJT	Supervisory Simulation	Lead./Sup. Rating
Dominance	(.88)							
Dependability	.13	(.82)						
Work Orientation	.60	.35	(.91)					
Supervisory Experience	.40	.02	.26	(.85)				
Supervisory Training	.27	.05	.21	.36	(.85)			
SJT	.18	.24	.19	.16	.16	(.75)		
Supervisory Simulation	.24	.10	.18	.13	.18	.21	(.72)	
Lead./Sup. Rating	.25	.10	.31	.26	.37	.20	.17	(.64)

Note. Sample size is 413.

Table 23

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 2 for Dominance

	Model A	Model B	Model C
Dominance->Experience Path	.47	.46	.46
Dominance->Training Path	--	.16	.16
Dominance->SJT Path	--	--	.18
Experience->Training Path	.43	.35	.35
Training->SJT Path	.21	.21	.15
Chi-square (df)	15.95 (3)	9.11 (2)	1.07 (1)
Probability of Perfect Fit	p = .001	p = .01	p = .30
Probability of Close Fit	p = .03	p = .09	p = .51
Population Discrepancy (90% Confidence Interval)	.03 (.01-.07)	.02 (.00-.05)	.00 (.00-.02)
RMSEA (90% Confidence Interval)	.10 (.06-.15)	.09 (.04-.16)	.01 (.00-.13)

Note. Sample size is 413.

Table 24 presents the LISREL results and fit indices for the models used to test Hypothesis 2 as it relates to Dependability. These results clearly indicate that Dependability has a substantial direct effect on SJT scores. Further, the paths from Dependability to supervisory experience and training are quite small and not significantly different from zero, suggesting that Dependability is only related to SJT scores and is not related to training and experience at all.

Table 25 presents the results for Hypothesis 2 and Work Orientation. These results are almost identical to the results for Dominance. The only difference is that the direct path from Work Orientation to supervisory experience is quite a bit smaller than the corresponding path for Dominance. Table 22 shows that Dominance has a moderately high correlation with Work Orientation (.60), so the similarity in results is not surprising.

Results Related to Hypothesis 3

Hypothesis 3 was also tested four times, once for each of the four individual differences measures of interest. For each individual differences measure, three models were tested, and Figure 3 summarizes the information included these three models. Model A included only the paths represented by solid lines in Figure 3. Model B was identical to Model A, except that it also included the path from the individual differences measure to the supervisory simulation. Model C was identical to Model A except that it also included a path from the individual difference measure to the Leading/Supervising ratings. Because the full model (with all of the paths represented in Figure 3) has no degrees of freedom, it was not tested.

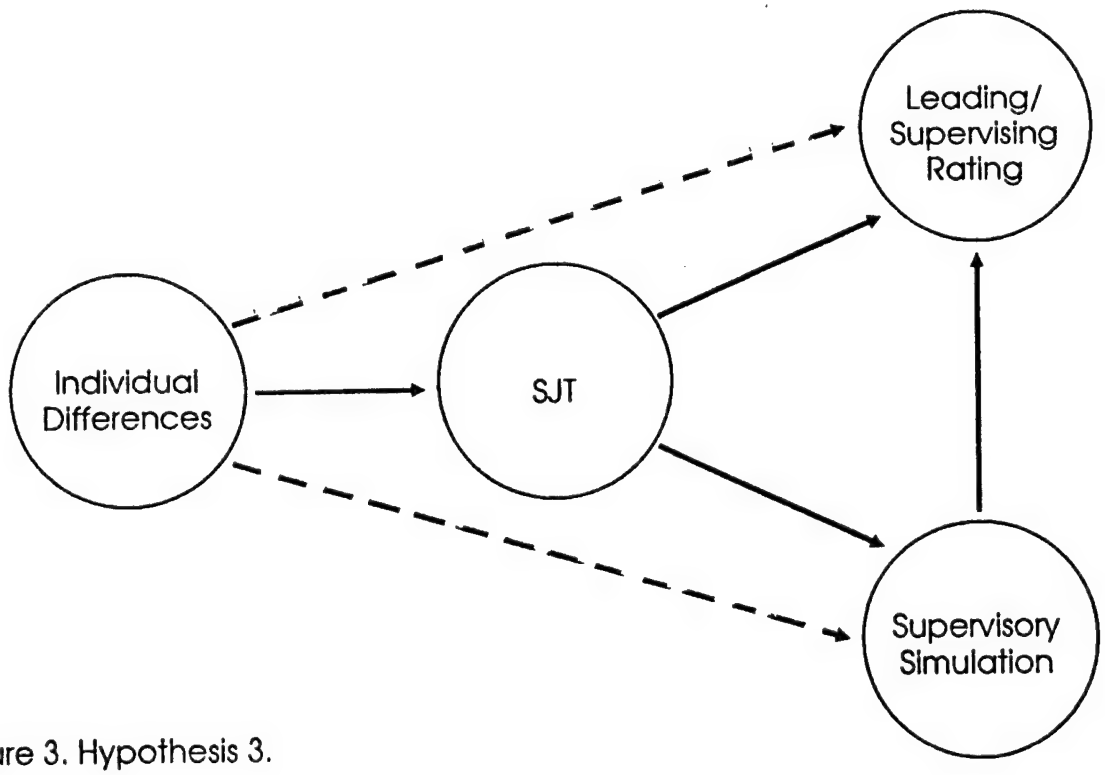

Figure 3. Hypothesis 3.

Table 24

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 2 for Dependability

	Model A	Model B	Model C
Dependability->Experience Path	.02	.02	.02
Dependability->Training Path	--	.07	.05
Dependability->SJT Path	--	--	.29
Experience->Training Path	.42	.42	.42
Training->SJT Path	.21	.21	.19
Chi-square (df)	28.01 (3)	26.62 (2)	4.36 (1)
Probability of Perfect Fit	$p < .001$	$p < .001$	$p = .04$
Probability of Close Fit	$p = .001$	$p < .001$	$p = .14$
Population Discrepancy (90% Confidence Interval)	.06 (.03-.11)	.06 (.03-.11)	.01 (.00-.03)
RMSEA (90% Confidence Interval)	.14 (.10-.19)	.17 (.11-.23)	.09 (.02-.18)

Note. Sample size is 413.

Table 25

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 2 for Work Orientation

	Model A	Model B	Model C
Work Orientation->Experience Path	.30	.30	.30
Work Orientation->Training Path	--	.14	.13
Work Orientation->SJT Path	--	--	.19
Experience->Training Path	.43	.38	.38
Training->SJT Path	.21	.21	.16
Chi-square (df)	17.81 (3)	11.74 (2)	1.93 (1)
Probability of Perfect Fit	$p < .001$	$p = .003$	$p = .17$
Probability of Close Fit	$p = .02$	$p = .04$	$p = .36$
Population Discrepancy (90% Confidence Interval)	.04 (.01-.08)	.02 (.01-.06)	.00 (.00-.02)
RMSEA (90% Confidence Interval)	.11 (.06-.16)	.11 (.06-.17)	.05 (.00-.15)

Note. Sample size is 413.

Table 26 shows the correlation matrix and reliability estimates that were used to test Hypothesis 3 as it relates to AFQT scores, and Table 27 presents the LISREL results and fit indices. Model B appears to have the best fit; even the estimate of RMSEA is .00 for this model. This indicates that the AFQT->Rating path is really not necessary to account for the observed correlations. Model A does not fit too badly either, indicating that the AFQT->Simulation path is not exceedingly important. These results suggests that relationships with SJT scores can completely account for the (small) correlation between AFQT scores and Leading/Supervising ratings and that SJT scores can at least partially account for the AFQT-simulation correlation.

Results related to Hypothesis 3 for Dominance are presented in Table 28. Model A does not fit very well at all. Models B and C fit somewhat better than Model A according to most of the fit indices, but the RMSEA results (which take into account the differences in the degrees of freedom) indicate that the fit is very similar for all three models. Taken together, these results suggest that Dominance has approximately equal, moderately sized direct effects on the SJT, simulations, and Leading/Supervising ratings, and that all of these direct effects are necessary to account for the observed correlations.

Table 29 presents the LISREL results and fit indices for Hypothesis 3 and Dependability. The fit for Model A is extremely good by all counts, and the path from Dependability to the SJT is substantial. It appears that Dependability has its only important direct effect on SJT scores. Paths to the simulations and Leading/Supervising ratings in Models B and C are very small and do not improve the fit appreciably.

Table 30 presents the results for Hypothesis 3 and Work Orientation. Again Work Orientation has a pattern of results very similar to the results for Dominance. The main differences are that the path from Work Orientation to Leading/Supervising ratings is substantially larger than the corresponding path for Dominance (.35 versus .26), and the path from Work Orientation to the simulations is somewhat smaller than the corresponding path for Dominance. These differences in the results for Work Orientation and Dominance are even more interesting in light of how highly correlated these two scales are. It appears that Work Orientation has a larger (positive) effect on supervisors' ratings, while Dominance is more strongly related to performance in the supervisory simulation exercises.

Results Related to Hypothesis 4

Three models were tested for Hypothesis 4, and Figure 4 summarizes the information included these three models. Again, Model A included only the paths represented by solid lines in the figure. Model B was identical to Model A except that it also included the training->simulation path, and Model C was identical to Model B except that it also included the training->ratings path. Table 31 presents the correlation matrix and reliability estimates that were used in testing these three models, and Table 32 shows the results. The fit for Model B is somewhat better than that for Model A, but both fit the data very poorly. Model C fits the data well, and is not rejected by the test of close fit. These results suggest that supervisory training does have

important, direct effects on both the supervisory simulation and the Leading/Supervising ratings. In addition, in Model C the training->ratings path is twice as large as the training->simulations path.

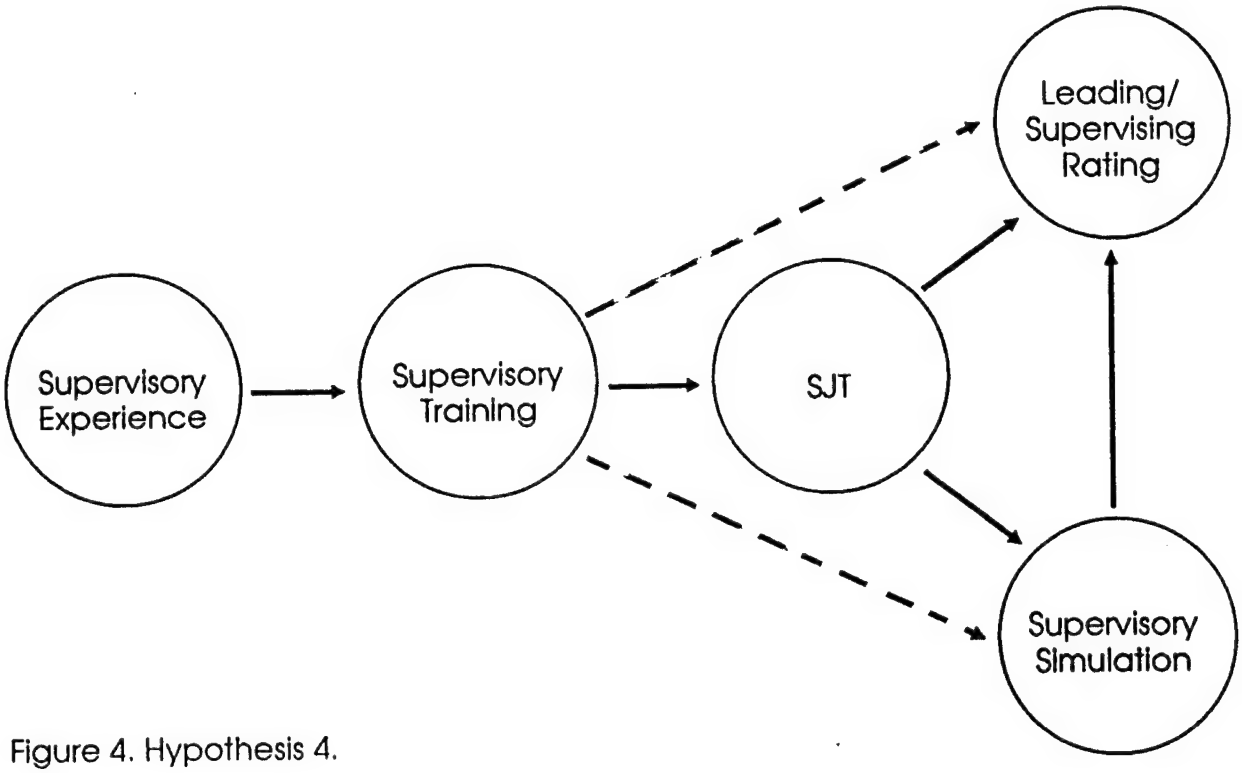

Figure 4. Hypothesis 4.

It is interesting to note that Model C fits the data quite well, even though it does not include direct paths from supervisory experience to the SJT, the simulations, or the Leading/Supervising ratings. Apparently all of the correlations between supervisory performance measures and supervisory experience can be accounted for by indirect effects through supervisory training. It also is interesting that in Model C the path from the simulations to the Leading/Supervising ratings is quite small, and it is not significantly different from zero. In retrospect, the Leading/Supervising ratings and the supervisory simulations can both be considered ratings of leadership performance; they are just based on different samples of behavior. In this context, the modeling results could be interpreted as indicating that the observed correlation between these two sets of ratings is primarily due to the fact that they have similar antecedents: supervisory training and knowledge.

One final note regarding the structural modeling analyses: all of the models involving frequency of supervisory responsibility (i.e., supervisory experience) were tested a second time replacing this supervisory experience

Table 26

Correlation Matrix Used in Testing Hypothesis 3 for AFQT Scores

	AFQT	SJT	Supervisory Simulation	Lead./Sup. Rating
AFQT	(.93)			
SJT	.30	(.75)		
Supervisory Simulation	.18	.19	(.72)	
Lead./Sup. Rating	.11	.20	.13	(.64)

Note. Sample size is 571.

variable with time as a supervisor. Results were virtually identical to those for supervisory experience and are therefore not presented here.

Conclusions Concerning SJT Relationships with Other Measures

The relationships of SJT scores with scores on other job performance measures are generally consistent with the interpretation of the SJT as a measure of supervisory job knowledge. The SJT is moderately correlated with other measures of supervisory performance (i.e., supervisory simulations and Leading/Supervising ratings), with measures of supervisory experience and training, and with promotion rate.

Results of the structural modeling analyses also support the construct validity of the SJT as a measures of supervisory job knowledge and provide additional information concerning the relationships of SJT scores with other measures. For example, results of the structural modeling analyses suggest that the relationship between supervisory experience and SJT scores can be accounted for by their relationships with supervisory training. In other words, these results suggest that more experienced soldiers are more frequently sent to supervisory training, and it is this training that is instrumental in raising SJT scores. Apparently the supervisory knowledge that is tapped by the SJT is more likely to be learned through relevant training than through on-the-job experience. The relationships between supervisory experience and scores on the other two supervisory performance measures (i.e., the simulation exercises and the Leading/Supervising ratings) can also be accounted for by supervisory training. In addition, supervisory training has substantial direct effects on all three measures of supervisory job performance.

Table 27

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 3 for AFQT

	Model A	Model B	Model C
AFQT->SJT Path	.37	.36	.37
AFQT->Simulation Path	--	.15	--
AFQT->Ratings Path	--	--	.03
SJT->Simulation Path	.27	.20	.27
SJT->Rating Path	.26	.25	.24
Simulation->Rating Path	.13	.13	.13
Chi-square (df)	7.48 (2)	.16 (1)	7.27 (1)
Probability of Perfect Fit	p = .02	p = .69	p = .007
Probability of Close Fit	p = .21	p = .84	p = .07
Population Discrepancy (90% Confidence Interval)	.01 (.00-.03)	.00 (.00-.01)	.01 (.00-.03)
RMSEA (90% Confidence Interval)	.07 (.02-.13)	.00 (.00-.08)	.11 (.04-.18)

Note. Sample size is 571.

Table 28

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 3 for Dominance

	Model A	Model B	Model C
Dominance->SJT Path	.25	.23	.24
Dominance->Simulation Path	--	.26	--
Dominance->Ratings Path	--	--	.26
SJT->Simulation Path	.31	.23	.31
SJT->Rating Path	.25	.24	.19
Simulation->Rating Path	.17	.21	.13
Chi-square (df)	30.57 (2)	13.30 (1)	16.09 (1)
Probability of Perfect Fit	$p < .001$	$p < .001$	$p < .001$
Probability of Close Fit	$p < .001$	$p = .004$	$p = .001$
Population Discrepancy (90% Confidence Interval)	.07 (.04-.12)	.03 (.01-.07)	.04 (.01-.08)
RMSEA (90% Confidence Interval)	.19 (.13-.25)	.17 (.10-.26)	.19 (.12-.28)

Note. Sample size is 413.

Table 29

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 3 for Dependability

	Model A	Model B	Model C
Dependability->SJT Path	.31	.30	.31
Dependability->Simulation Path	--	.05	--
Dependability->Ratings Path	--	--	.05
SJT->Simulation Path	.29	.27	.29
SJT->Rating Path	.24	.24	.22
Simulation->Rating Path	.18	.18	.18
Chi-square (df)	1.02 (2)	.55 (1)	.43 (1)
Probability of Perfect Fit	p = .60	p = .46	p = .51
Probability of Close Fit	p = .82	p = .65	p = .69
Population Discrepancy (90% Confidence Interval)	.00 (.00-.01)	.00 (.00-.01)	.00 (.00-.01)
RMSEA (90% Confidence Interval)	.00 (.00-.05)	.00 (.00-.12)	.00 (.00-.11)

Note. Sample size is 413.

Table 30

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 3 for Work Orientation

	Model A	Model B	Model C
Work Orientation->SJT Path	.26	.24	.24
Work Orientation->Simulation Path	--	.18	--
Work Orientation->Ratings Path	--	--	.34
SJT->Simulation Path	.30	.24	.30
SJT->Rating Path	.26	.25	.17
Simulation->Rating Path	.17	.20	.13
Chi-square (df)	35.68 (2)	26.98 (1)	.7.61 (1)
Probability of Perfect Fit	$p < .001$	$p < .001$	$p = .006$
Probability of Close Fit	$p < .001$	$p < .001$	$p = .04$
Population Discrepancy (90% Confidence Interval)	.08 (.04-.14)	.06 (.03-.11)	.02 (.00-.05)
RMSEA (90% Confidence Interval)	.20 (.15-.26)	.25 (.18-.34)	.12 (.06-.22)

Note. Sample size is 413.

Table 31

Correlation Matrix Used in Testing Hypothesis 4

	Supervisory Experience	Supervisory Training	SJT	Supervisory Simulation	Lead./Sup. Rating
Supervisory Experience	(.85)				
Supervisory Training	.38	(.85)			
SJT	.17	.18	(.75)		
Supervisory Simulation	.13	.20	.20	(.72)	
Lead./Sup. Rating	.26	.34	.22	.16	(.64)

Note. Sample size is 698.

Correlations between SJT scores and scores on selected temperament and cognitive ability measures are also consistent with the interpretation of the SJT as a measure of supervisory job knowledge. The correlation between SJT scores and AFQT scores is quite high but at about the same level as the correlation between the AFQT and technical job knowledge test scores. The SJT is moderately correlated with Dominance, Dependability, and Work Orientation.

Results of the structural modeling analyses indicate that much of the effect of Dominance and Work Orientation on SJT scores can be accounted for by indirect effects through amount of supervisory experience and training. This means that at least part of the relationship between Dominance and Work Orientation and the SJT could be due to more dominant, hard working soldiers obtaining more supervisory training and experience and this training and experience in turn leading to higher SJT scores. However, both general cognitive ability (AFQT) and Dependability clearly have moderately large, direct effects on SJT scores that are not mediated by supervisory training or experience. Results for the item type subscales suggest that the Dependability-SJT correlation is mostly due to more dependable soldiers recognizing when disciplining is appropriate. SJT scores also appear to mediate the relationships between AFQT scores and scores on the other supervisory performance measures, and this is what would be expected for a measure of supervisory job knowledge. On the other hand, Dominance and Work Orientation have significant, direct effects on all three measures of supervisory performance.

Table 32

Path Coefficients, Chi-Square, Fit Indices, and Residuals for Models Related to Hypothesis 4

	Model A	Model B	Model C
Experience->Training Path	.45	.45	.45
Training->SJT Path	.27	.22	.22
Training->Simulation Path	--	.20	.20
Training->Ratings Path	--	--	.39
SJT->Simulation Path	.29	.23	.23
SJT->Ratings Path	.32	.28	.22
Simulation->Ratings Path	.15	.16	.08
Chi-square (df)	94.10 (5)	75.01 (4)	14.11 (3)
Probability of Perfect Fit	$p < .001$	$p < .001$	$p = .003$
Probability of Close Fit	$p < .001$	$p < .001$	$p = .13$
Population Discrepancy (90% Confidence Interval)	.13 (.09-.18)	.10 (.07-.15)	.02 (.00-.04)
RMSEA (90% Confidence Interval)	.16 (.13-.19)	.16 (.13-.19)	.07 (.04-.11)

Note. Sample size is 698.

Relationships between scores on the subscales based on relevance of special training (Training Needed, Training May or May Not be Needed, Training Not Needed) and other measures provide only modest support for the notion that some SJT items are more like ability measures while others are more achievement related. In general, all three subscales have similar patterns of correlations with other measures, although the Training Needed subscale has somewhat higher correlations with supervisory training and with measures of supervisory performance.

Correlations between the item type subscales and the other measures suggest that these subscales do tap somewhat different aspects of supervisory knowledge. For example, the supervisory simulations are significantly correlated with all of the item type subscales that involve dealing with subordinates, but they are not significantly correlated with the Chain of Command subscale. Searching for Reasons has the highest correlation with AFQT scores, while Dependability most highly correlated with the Disciplining as Appropriate subscale. These results are consistent with the notion that the supervisory knowledge that is measured by the SJT is actually several somewhat distinct but intercorrelated aspects of supervisory job knowledge. Many of the item type subscales are short and have low internal consistency reliabilities, but these subscales provide a certain amount of information concerning the nature of the specific knowledges measured by the SJT.

CONCLUSIONS AND RECOMMENDATIONS

The results of the basic SJT data analyses indicate that this test was appropriately difficult for the CVII sample. There was initially some concern that the correct answers on the SJT would be too obvious, so this is encouraging. Internal consistency reliabilities and item-total correlations are quite high. Investigations of the dimensionality of the SJT demonstrated that the SJT response alternatives do, in fact, describe a wide variety of supervisory behaviors, and these behaviors cover most of the range of supervisory tasks identified in the earlier job analysis. Results of analyses for the "item type" subscales indicate that the SJT may actually measure several related aspects of supervisory job knowledge. The lack of clear results for the subscales based on the relevance of special training suggests that all of the SJT items are at approximately the same location along the aptitude/achievement continuum. Or, another possible explanation is that for a fair number of SJT items some soldiers know the correct answer based on general life experiences (i.e., the item measures an aptitude) while other soldiers must learn what is more effective through relevant training or experience (i.e., the item measures achievement).

The correlations between SJT scores and scores on other job performance measures, cognitive ability, and selected temperament measures provide a good deal of support for the construct validity of the SJT as a measure of supervisory job knowledge. Results of the structural modeling analyses show that the SJT has both the direct and indirect relationships with other measures that would be expected for a test of supervisory job knowledge. For example, these results suggest that the SJT mediates the relationships between general mental ability (AFQT) and the supervisory simulations and Leading/Supervising ratings. This supports the notion that the SJT measures supervisory job knowledge because soldiers have to know what to do before they can do it effectively, so general mental ability would be expected to have its effect on supervisory performance through this learning process. Soldiers with more supervisory training also obtained significantly higher SJT scores, indicating that the knowledges measured by the SJT are, to some extent, learned through relevant training.

Based on the results of the present research and other available research, the SJT can best be interpreted as a measure of supervisory job knowledge. Although investigations of the dimensionality of the SJT revealed several somewhat different aspects of the supervisory knowledge that is measured, the SJT subscores have some psychometric problems, and the SJT Total Score (M-L Effectiveness) provides the best summary of the information contained in the SJT.

The SJT items and the mean effectiveness ratings of the SJT response alternatives from the Sergeants Major Academy furnish an opportunity to explore what effective supervisory practices in the Army actually involve. The investigations of the effectiveness of various sources of power in the present research provide preliminary information. Mean effectiveness ratings are available for the response alternatives for all 180 items in the SJT developmental

work, and these data could be used to conduct further research as to the nature of effective supervisory practices in the Army.

Because the CVII data analysis results indicated that the SJT was a promising measure of supervisory performance, this test was lengthened for the next phase of the Career Forces project by adding fourteen new items from the original pool of 180 items. The 35-item SJT was very difficult for the CVII sample, so an effort was made to add relatively easy items to the test. In addition, correlations of scores on existing SJT items with scores on other supervisory performance measures were computed for the CVII sample, and an effort was made to include new items that had content similar to the content of existing SJT items that had both meaningful correlations with other measures and high item-total correlations with the SJT itself. The resulting 49-item SJT has been administered to the Project A/Career Forces Longitudinal Validation second-tour (LVII) sample, along with the other second-tour job performance measures. These data will provide an opportunity to further delineate the exact nature of the construct measured by the SJT. For example, data from this longer version of the SJT will likely lead to the identification of longer, more reliable SJT subscales. These subscales can be used in the modeling of second-tour soldier performance for the LVII sample, and may aid in understanding the aspects of soldier performance related to leadership and supervision.

REFERENCES

Alderman, D. L., Evans, F. R., & Wilder, G. (1981). The validity of written simulation exercises for assessing clinical skills in legal education. Educational and Psychological Measurement, 41, 1115-1126.

Angoff, W. H., & Johnson, E. G. (1988). A study of the differential impact of curriculum on aptitude test scores (Research Report No. RR-88-46). Princeton, NJ: Educational Testing Service.

Borman, W. C., Hanson, M. A., Oppler, S. H., Pulakos, E. D., & White, L. A. (in preparation). The role of early supervisory experience in supervisor performance. Journal of Applied Psychology.

Browne, M. W., & Cudeck, R. (in preparation). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), Testing structural equation models. Beverly Hills, CA: Sage.

Bruce, M. M., & Learner, D. B. (1958). A supervisory practices test. Psychology, 11, 207-216.

Brull, H. P. (1981). Written simulations: The state of the art, A literature review. Minneapolis, MN: Personnel Decisions, Inc.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986). Development and field test of Project A task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A182 645)

Campbell, J. P. (Ed.) (1991). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1988 fiscal year (ARI Research Note 91-34). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A233 750)

Campbell, J. P. (Ed.) (1989). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1987 fiscal year (ARI Technical Report 862). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A219 046)

Campbell, J. P., & Zook, L. (Eds.) (1990). Building and retaining the career force: New procedures for accessing and assigning Army enlisted personnel: Annual report, 1990 fiscal year (ARI Technical Report 952). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A252 675)

Campbell, J. P., & Zook, L. (Eds.) (in preparation). Building and retaining the career force: New procedures for accessing and assigning Army enlisted personnel: Annual report, 1991 fiscal year (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J. P., & Zook, L. (Eds.) (1991). Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A (ARI Research Report 1597). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A242 921)

Cudeck, R. (1989). Analyses of correlation matrices using covariance structure models. Psychological Bulletin, 105, 317-327.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

French, J.R.P., & Raven, B. (1959). The bases of social power. In D. Cartwright, & A. Zander (Eds.), Group dynamics (pp. 150-167). New York: Harper & Row.

Gunning, R. (1952). The technique of clear writing. New York: McGraw-Hill Book Company.

Hanson, M. A., & Borman, W. C. (1990, November). A situational judgment test of supervisory knowledge in the U.S. Army. Paper presented at the 32nd Annual Conference of the Military Testing Association, Orange Beach, AL.

Hanson, M. A., & Borman, W. C. (1989, April). Development and construct validation of a Situational Judgment Test as a job performance measure for first line supervisors. In W. C. Borman (Chair), Evaluating "Practical IQ": Measurement issues and research applications in personnel selection and performance assessment. Symposium conducted at the 4th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta.

Hough, L. M., Barge, B. N., & Kamp, J. D. (1987). Non-cognitive measures: Pilot testing. In N. G. Peterson (Ed.), Development and field test of the Trial Battery for Project A (ARI Technical Report 739), pp. 7-1 through 7-8. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A184 575)

Humphreys, L. L. (1974). The misleading distinction between aptitude and achievement tests. In D. R. Green (Ed.), The aptitude-achievement distinction (pp. 262-274). Monterey, CA: CTB/McGraw-Hill.

Joreskog, K. G., & Sorbom, D. G. (1981). LISREL VI. Mooresville, IN: Scientific Software, Inc.

Kipnis, D., Schmidt, S. M., & Wilkinson, I. (1980). Interorganizational influence tactics: Explorations of getting one's way. Journal of Applied Psychology, 65, 440-452.

Mandell, M. M. (1950). The administrative judgment test. Journal of Applied Psychology, 34, 145-147.

McCormick, B. K., Dunlap, W. P., Kennedy, R. S., & Jones, M. B. (1983). The effects of practice on the Armed Services Vocational Aptitude Battery (ARI Technical Report 602). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A148 314)

McGuire, C. H., & Babbott, D. (1976). Simulations techniques in the measurement of problem solving skills. Journal of Educational Measurement, 4, 1-10.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low fidelity simulation. Journal of Applied Psychology, 75, 640-647.

Mowry, H. W. (1964). Leadership evaluation and development scale casebook. Los Angeles, CA: Psychological Services, Inc.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

Mulder, M., de Jong, R. D., Koppelaar, L., & Verhage, J. (1986). Power, situation, and leaders' effectiveness: An organizational field study. Journal of Applied Psychology, 71, 566-570.

Murphy, K. R. (1984). Armed Services Vocational Aptitude Battery. In. D. J. Keiser, & R. C. Sweetland (Eds.), Test Critiques. Kansas City, MO: Test Corporation of America.

Podsakoff, P. M., & Schriesheim, C. A. (1985). Field studies of French and Raven's bases of power: Critique, reanalysis, and suggestions for future research. Psychological Bulletin, 97, 387-411.

Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (1987). The development of administrative measures as indicators of soldier effectiveness (ARI Technical Report 754). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A191 232)

Rosen, N. A. (1961). How supervise? -- 1943-1960. Personnel Psychology, 14, 87-99.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximal performance. Journal of Applied Psychology, 73, 482-486.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Smith, I. L. (1983). Use of written simulations in credentialing programs. Professional Practice of Psychology, 4, 21-50.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 39, 149-155.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Tenopyr, M. L. (1969). The comparative validity of selected leadership scales relative to success in production management. Personnel Psychology, 22, 77-85.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real world pursuits: The role of tacit knowledge. Journal of Personality and Social Psychology, 49, 436-458.

Yukl, G., & Falbe, C. (1990). Influence tactics and objectives in upward, downward, and lateral relations. Journal of Applied Psychology, 75, 132-140.

Yukl, G., & Falbe, C. M. (1991). Importance of different power sources in downward and lateral relations. Journal of Applied Psychology, 76, 416-423.

Appendix A: Item/Response Alternative Rating Task

SJT Rating Task: Instructions

The Situational Judgment Test (SJT) is a multiple choice test designed to measure supervisory skill. SJT items describe difficult supervisory situations that a first-line supervisor in the Army might encounter. There are between three and five response alternatives for each item, and these response alternatives describe possible actions that the supervisor might take in that situation.

This rating task involves making several ratings concerning the content of SJT item stems and response alternatives. The first four rating categories (A through D) involve rating the content of the item stem (i.e., the situation). The last rating category (H) involves rating the content of the response alternatives. These two types of ratings will be used, in combination, to understand why certain response alternatives are more effective than others. (For example, it may be the case that when the situation or item stem involves dealing with a very mature, responsible subordinate with a minor disciplinary problem that response alternatives involving punishment are always quite ineffective.)

This rating task also involves making several ratings concerning the item as a whole. For the middle three rating categories (E through G) you will be asked to consider each item's content in combination with the effectiveness values (which range from 1 to 7) for each response alternative. You will essentially use the content of the item stem and the response alternatives to interpret ***why*** certain response alternatives are more effective than others. In doing this, the key will be the ***relative*** effectiveness of the various alternatives. For example, one alternative may be only average (e.g., 3.5 on the 7-point scale) while another is extremely effective. What is it that makes the first alternative so much more effective than the second? In many cases the absolute level of effectiveness is also interesting. In the previous example you might ask yourself, why is the second alternative average as opposed to extremely ineffective?

In making these ratings that involve the effectiveness values, consider the effectiveness of *all* of the response alternatives for each item. For some items, the second most effective response alternative is almost as effective as the most effective response alternative. For other items the most effective response alternative is quite a bit more effective than any other response alternative. In addition, some items have response alternatives that are extremely ineffective (e.g., a rating of about 1 or 2), while for other items the least effective response alternative is similar in effectiveness to one or more other response alternatives. Consider the effectiveness of *all* of the response alternatives for a particular item when you make your ratings. For example, if an item has two very effective response alternatives and two very ineffective response alternatives, think about what it is about *both* of the relatively effective response alternatives that makes them more effective than the two ineffective alternatives.

The effectiveness values come from previous workshops in which senior NCOs at the U.S. Army Sergeants Major Academy rated the effectiveness of each response alternative for each SJT item. The effectiveness values for each response alternative appear in the left margin of the rating sheets. These ratings were made using the 1 to 7 scale shown below:

1	2	3	4	5	6	7
Very Ineffective						Very Effective

To complete the present rating task, you will review the SJT items, along with the effectiveness value for each response alternative (from the Sergeants Major Academy) and then make several ratings. Definitions of the rating categories are provided on a separate sheet. Please read the definition of each category and the various anchors for each category carefully before making your ratings. Each of the ratings should be made independently of each other. It would probably be easiest to rate the first four dimensions at the same time, then go back through the items and rate the remaining dimensions.

Some of the rating categories require additional clarification -

For rating category A (Objective), levels 2 and 5, please include the small letters in your ratings. For example, if you think the item stem involves a minor disciplinary problem you would rate that item "5a" on category A. If you think the item has only one objective, please put a "0" in the space provided for the secondary objective.

For rating category D (Maturity or Responsibility Level of Target Person(s)) we are referring to how dependable the target person is or the typical level of responsibility or maturity that is exhibited by the target person. If this category seems difficult to rate for some items because it is too multidimensional, please make a note of which items were difficult to rate and why.

For rating categories E through G, the main point is that you should use the effectiveness value (i.e., the number in the left margin) for each response alternative ***and*** the content of these alternatives to make your ratings. For the category E, Relevance of Special Training/Knowledge, you will be indicating the extent to which you believe a person would need Army training or Army supervisory training to know which responses are more effective and which are less effective. By relevant training/knowledge we do ***not*** mean an understanding of the terminology used in these items (e.g., "Article 15"). As you make these ratings, you might assume that people taking the test would be provided with a list of definitions for these terms. Also, if you do not know the meaning of any of these terms (especially the abbreviations) please check with

me (or with someone who is particularly familiar with Army terminology). If you don't think an item has a secondary type, please put a "0" in the space provided.

SJT Rating Task: Definitions of Rating Categories and Anchors

Rating Categories A through F

When making your ratings for the first four categories, focus on the content of the item stem (i.e., the description of the situation). You can use the response alternatives to help clarify the situation, but your ratings should focus, as much as possible, on the situation itself.

Rating Category A: Objective (primary and secondary)

What is the respondent's objective or goal in the situation described in this item? (i.e., What are they trying to accomplish, either immediately or in the long run?) If there appears to be more than one objective or goal for an item, choose the one that seems most related to the effectiveness of the response alternatives as the primary objective (e.g., If all of the response alternatives for an item are similar in terms of how effectively tasks or projects are assigned, but some response alternatives better reward good performance you would assign that item a "3".) Then, list any other objective as the secondary objective.

1. Assign task(s) or project(s)

2. Improve substandard subordinate performance
   a. chronic/long term performance problem
   b. relatively short term or recently developed performance problem
   c. immediate performance problem

3. Reward good subordinate performance

4. Solve personal problems that may interfere with work performance

5. Discourage or prevent breaches of discipline or other serious misconduct
   a. Minor disciplinary problems
   b. Serious misconduct

6. Obtain a change in plans (other than improved performance)

7. Provide subordinate(s) with encouragement or support

8. Other (please describe briefly)

0. Objective is not specified or not clear

Rating Category B: Direction of Interaction/Target Person(s)

Toward whom are the respondent's actions directed? This is not necessarily the person that they would talk to, but it is the person their actions are intended to affect. Who is the target person (i.e., the person they are trying to influence)? This is often one or more subordinates, but it is sometimes a supervisor or peer.

0. Not applicable
1. Upward (e.g., his/her supervisor)
2. Lateral (e.g., a peer)
3. Downward (i.e., one or more subordinates)
4. Mixed (e.g., respondent could choose to direct their actions toward a supervisor *or* subordinate in response to this situation)

Rating Category C: Performance of Target Person(s)

What is the current performance level of the person the respondent is trying to influence in this situation. If more than one of these apply to a situation, please list all that apply. List the aspects of performance that are most relevant to the situation ***first***. (For example, if a subordinate is described as a poor performer, but a minor disciplinary infraction appears to be the central aspect of the situation you might rate that item 8/4.)

1. Outstanding
2. Very good/above average
3. Adequate/good
4. Poor
5. Very poor
6. Chronic problem(s) (e.g., performance problem)
7. Declining performance
8. Minor disciplinary problem
9. Serious disciplinary problem
10. Personal problem (where performance may or may not be affected)
0. Not applicable or not mentioned

Rating Category D: Maturity or Responsibility Level of Target Person(s)

What is the typical maturity or responsibility level of the target person(s)? This is not necessarily related to their current performance (e.g., the item might describe the target person's typical performance and their current performance separately - in this case rate the maturity/responsibility level related to their ***typical*** performance). However, if the current performance gives a clear indication of the target person's typical maturity or responsibility and there is no other information provided then use their current performance to make this rating.

1. Extremely mature and/or responsible
2. Very mature and/or responsible
3. Reasonably mature and/or responsible
4. Not particularly mature or responsible
5. Irresponsible or immature
6. Very irresponsible or immature
7. Extremely irresponsible or immature
0. Not applicable or not mentioned

Rating Categories E through G

When making your ratings for the next three categories, consider the content of the item stem, the content of the response alternatives, the effectiveness of the response alternatives, and the differences among the response alternatives that appear to account for their different effectiveness values. In other words, what is it that this item is really getting at or measuring?

Rating Category E: Relevance of Special Training/Knowledge

To what extent does identifying the more effective responses to this item appear to require special training or knowledge (e.g., familiarity with military supervisory procedures). In other words, when you look at the effectiveness of the response alternatives for an item, do these values make sense to you, or does it appear that these effectiveness values reflect knowledge that you do not have (i.e., that might be trained in the Army). When making this rating, assume that everyone taking the test understands the meaning of the Army terminology used: for example "UCMJ action," "ACS," "CQ," "class A uniform," etc.

1. This item ***definitely*** requires special training or knowledge (i.e., the

effectiveness values for the response alternatives appear counter-intuitive).

2. This item ***probably*** requires special training or knowledge.

3. This item may or may not require special training or knowledge.

4. This item ***probably doesn't*** require special training or knowledge.

5. This item ***clearly doesn't*** requires special training or knowledge (i.e., the effectiveness values for the response alternatives make perfect sense).

0. This item cannot be rated in terms of special training or knowledge.

Rating Category F: Primary Item Type

The item "types" listed here try to capture the essence of the situations. How is it that the more effective responses differ from the less effective responses? If more than one of the "types" listed here apply to an item, choose the one that is most important in determining the effectiveness of responses for that item. Then, for each response alternative assign a "+" if it is particularly effective relevant to that type, a "-" if it is particularly ineffective an "=" if it is not particularly effective or ineffective, and a "?" if that aspect of the item is not relevant to that particular alternative.

1. Search for underlying personal problems that are affecting work when appropriate

2. Search for underlying causes or reasons for problems (other than personal problems) and/or gather information relevant to possible underlying causes

3. Avoid inappropriately harsh discipline

4. Discipline or take other severe administrative actions when severity of the misconduct or chronic nature of problem dictates

5. Clarify performance standards or the consequences of target person's actions when necessary

6. Resist being taken in by subordinates' stories (i.e., not "putting up with any crap")

7. Assign tasks effectively and/or according to established procedures

8. Acknowledge or emphasize the positive (e.g., past good performance)

9. Work through the proper chain of command as appropriate

10. Provide subordinates with needed support and/or encouragement

11. Ensure that subordinates obtain appropriate rewards

0. None of the types listed above apply to this item

Rating Category G: Secondary Item Type(s)

If more than one of the "types" listed above applies to an item, record the second type in this space. For example, if a situation involves avoiding inappropriately harsh discipline, but primarily requires the respondent to search for an underlying personal problem, you would record avoiding inappropriately harsh discipline as the secondary item type. If there are several secondary item types, record them all in this space in the order of importance.

Rating Category H

When making your ratings for the final category, focus on the ***content*** of each response alternative. For this last category you will make a rating for ***each*** response alternative.

Rating Category H: Source of Power

Power refers to an agent's capacity to influence a target person's behavior. What is the ***primary*** source of power that a respondent is ***attempting*** to use to influence the target person's behavior in each response alternative? If a response alternative clearly involves more than one source of power, list them in the order of importance. In making these ratings, keep in mind the respondent's objective or goal in the situation. What is it that they are trying to get the person(s) to do?

Position as a source of power:

1. Legitimate power - involves the formal authority associated with the agent's position. They simply tell people what to do and expect them to do it.

2. Reward power - influencing the target person(s) by offering or providing rewards (e.g., awards, promotions, a better work schedule, etc.).

3. Coercive power - influencing the target person(s) by threatening to punish them or to withhold desired rewards or by actually administering punishment or withholding rewards.

4. Information power - influencing the target person(s) by obtaining or providing certain information.

Personal attributes as a source of power:

5. Expert power - influencing the target person(s) by providing (or threatening to withhold) advice or expert assistance.

6. Persuasive power - influencing the target person(s) by making logical arguments or presenting evidence that appears credible.

7. Not clear which source of power is being used.